THE UNIVERSITY OF CHICAGO

RECOVERING STRUCTURE FROM OBSERVATIONAL DATA: REPRESENTATION, SIMULATION, AND STABILITY

A DISSERTATION SUBMITTED TO
THE FACULTY OF THE DIVISION OF THE PHYSICAL SCIENCES
IN CANDIDACY FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

DEPARTMENT OF STATISTICS

BY

KULUNU DHARMAKEERTHI

CHICAGO, ILLINOIS

AUGUST 2026

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

I begin with my deepest gratitude to my advisor, Professor Tengyuan Liang. Tengyuan has been an ideal model of what a scholar can be: mathematically broad, intellectually fearless, and able to distill a problem to its essential structure. What has shaped me most, however, is his taste for research. In our conversations, the question was never only whether a problem could be solved, but whether it was interesting, whether it clarified something, and whether it might be useful to others. He taught me to value practical and scientific purpose without losing sight of mathematical beauty. He has given me the freedom to find my own questions, while offering guidance whenever the path became unclear. At difficult moments, his patience, perspective, and confidence in me often helped restore my own. I am grateful not only for what he taught me about research, but for the steadiness, generosity, and care with which he guided me through graduate school. I could not have asked for a better advisor, and I will carry the example of his scholarship and mentorship with me throughout my career.

I am also grateful to the members of my committee, Professors Rina Barber and Cong Ma. I have had the privilege of learning from both of them, either in the classroom or through their research groups early in my Ph.D. They have shaped my education in ways that reach well beyond the contents of this thesis. I thank them for their time, their guidance, and for reading this work with such care.

Graduate school would have been immeasurably poorer without the friends and collaborators with whom I shared so many conversations, frustrations, and discoveries. I am grateful to many. Sean O'Hagan, Subhodh Kotekal, Omar Ghattas, Aabesh Bhattacharya, Soumyabrata Kundu, Annie Xie, Kiho Park, Raphael Rossellini, YoonHaeng Hur and others. Their work ethic, and approach to research have continually inspired me. Many of our conversations began with mathematics or statistics, but they often became something broader: a reminder that intellectual life is sustained by friendship, curiosity, humor, and the pleasure of thinking together. Their passion for their own work was infectious, and I learned a great

deal simply by being around them.

I owe a special debt to YoonHaeng Hur, who was both a coauthor and a mentor to me early in my Ph.D. At a stage when much of research still felt opaque, his guidance helped make the process more intelligible and less solitary. I am grateful not only for what we worked on together, but for the example he set: careful, generous, and kind.

I also want to thank my friends outside the university. Luis Ohlendorf, Shayden Chrisostom, Alex Taylor-Brown, Isuru Gunasekara, Kulindu Hirimuthugoda and many others back home. They have known me long before this journey began, and for that reason their friendship has been a particular source of comfort. They reminded me, often at exactly the right moments, that there was a world outside graduate school. They made me laugh, kept me sane, and gave me the kind of companionship that no professional accomplishment can replace.

I am also deeply grateful to my family: my cousins, aunts, uncles, and grandparents. Though we have often been separated by distance, I have always felt their love, encouragement, and support. Their belief in me has meant more than I can say. I hope, one day, to support them with even a fraction of the generosity with which they have supported me.

My greatest and oldest debt is to my parents. Their sacrifices are the reason I am here today. Without their love, support, and belief in me, none of this would have been possible. They have given me more than I can ever repay: the opportunity to pursue an education, the confidence to continue when things were difficult, and the example of perseverance, generosity, and quiet strength. I am forever indebted to them, and they continue to inspire me in ways I am only beginning to understand.

I am also grateful to my brother, Dasula, whose determination, curiosity, and passion have been a continuous source of inspiration. I admire the energy with which he approaches the world, and I know he will do remarkable things. His presence has made my life brighter, warmer, and more joyful.

Finally, I thank Kyra Kucirka. Your support has carried me through this process in ways

that are difficult to put into words. Your kindness, and generosity are a daily example to me. This thesis would not have been possible without you, and it means far more because you were beside me through it.

ABSTRACT

Modern statistical learning increasingly begins from data that were observed rather than designed. Such data are records of an uncontrolled process. They may contain rich information about the environment that produced them, but they do not by themselves say what should be learned, which variation is useful, or which parts of the data-generating process can be trusted for inference. The central problem is therefore interpretive as well as computational.

We develop this view through three studies of observed environments. The first analyzes diffusion-based generative modeling as a way to represent an observed data distribution. Empirically, these models have been shown to capture features of an environment that are difficult to describe parametrically: dependence, heterogeneity, multimodality, and geometric structure. We ask what makes this possible statistically, and how the denoising operations used by diffusion models recover structure in the underlying distribution. The second study asks what becomes possible once observed environments can be accessed by simulation. Using learned conditional generators, we develop a simulation-based approach to conditional moment restrictions, with nonlinear instrumental variables as the leading example. The third study considers what happens when the observed environment changes. It examines prediction under concept shift, where unobserved environment-specific variation can make a predictor learned in one environment fail to transfer to another. Together, these studies examine how modern learning procedures can be used to recover, simulate, and stabilize structure in observational data.

CHAPTER 1

INTRODUCTION

Statistical learning begins with observations, but observations are rarely made for the statistician's convenience. In an experiment, the statistical question is built into the design: the researcher controls assignment, intervention, or sampling so that the data-generating process is aligned with the question of interest. Such designs, however, are scarce, costly, and often infeasible. Observational records are far more abundant. They are produced continuously by scientific, economic, administrative, and social activity, often before any formal statistical question has been posed. Their promise and difficulty are inseparable: more of the world is measured and recorded than ever before, but the statistician does not design the processes that produces these data.

Because observational data are given rather than designed, statistical analysis begins with interpretation. An observation is evidence of a data-generating process; it is not, on its own, an explanation of that process. The mechanism that produced the data is only partially observed, and the data do not specify the statistical role they should play. The same observations may therefore admit many descriptions. Which description is appropriate depends on the question, the assumptions, and the target object. The first difficulty is thus interpretive: one must identify which features of the observed distribution are informative for the object of interest.

This thesis is organized around two questions at the center of statistical learning from observational data:

What is to be recovered from the observed environment?

What structure in the observations makes that recovery possible?

Together, they impose a discipline on the analysis of observational data: one must first state what is to be recovered, and then articulate the structure in the observed distribution

that permits its recovery.

This thesis applies this perspective to three modern statistical problems in which the object of interest is not directly observed: learning complex data distributions, recovering obscured structural relations, and extracting predictive structure that remains stable across observed environments.

Across these settings, a central theme is the relation between modern machine learning and classical statistical reasoning. Machine learning supplies flexible procedures for approximation, simulation, and representation. Statistical reasoning asks when these procedures recover a well-defined object and what features of the observed data enable identifying that object. The thesis therefore treats modern learning methods not only as computational tools, but as statistical procedures whose meaning depends on the structure of the observational environment.

1.1 Representing Observed Environments

The first problem considered in the thesis is the representation of an observed environment. Suppose

$$X_1, \ldots, X_n \sim \mu,$$

where μ is an unknown probability law on $\mathbb{R}^d$. The observations are measurements taken jointly from an environment. Their coordinates may have names, meanings, dependencies, and latent organization inherited from the process that generated them. A statistical representation of this environment should therefore capture more than marginal summaries or isolated features. It should encode how observations are organized under the joint law.

Generation provides a stringent test of such a representation. To generate new observations from μ, a fitted object must recover enough of the dependence, geometry, heterogeneity, and concentration structure of the distribution to produce plausible new draws. In this sense, sampling is a form of representation: a model that can generate from the observed law has

learned a reusable statistical description of the environment that produced the data.

This is a demanding objective. In a parametric model, one restricts μ to a family $\{\mu_\theta : \theta \in \Theta\}$ and estimates θ. In many modern applications this is an inadequate starting point. The law μ may be high-dimensional, multimodal, non-log-concave, or concentrated near structure that is difficult to describe analytically. Modern generative modeling asks whether such a law can be represented directly through a sampling procedure.

Chapter 3 studies this question through diffusion-based sampling. Diffusion models approach generation indirectly. They construct a forward noising process that gradually transforms μ into a simpler reference law. At the level of probability measures, this gives a forward path

$$\mu = \mu_0 \xrightarrow{f_1^\mu} \mu_\eta \xrightarrow{f_2^\mu} \cdots \xrightarrow{f_K^\mu} \mu_{K\eta} \approx \nu,$$

where ν is an easy-to-sample reference law. The forward process is tractable because it smooths the original distribution.

Sampling requires moving in the opposite direction. If one could start exactly from $\mu_{K\eta}$ and apply the correct backward maps, the original law could be recovered:

$$\mu = \mu_0 \xleftarrow{b_1^\mu} \mu_\eta \xleftarrow{b_2^\mu} \cdots \xleftarrow{b_K^\mu} \mu_{K\eta}.$$

The statistical issue is hidden in this idealized display. Practical sampling starts from a simple distribution ν, rather than from the exact terminal law $\mu_{K\eta}$. The actual backward path is better written as

$$\bar{\nu}_0 \xleftarrow{b_1^\mu} \bar{\nu}_\eta \xleftarrow{b_2^\mu} \cdots \xleftarrow{b_K^\mu} \bar{\nu}_{K\eta} := \nu.$$

The question is whether $\bar{\nu}_0$ is close to μ.

This is a sensitivity problem. The forward diffusion may map many different initial laws close to the same reference law. Proximity of ν to $\mu_{K\eta}$ is therefore only part of the story.

One must understand how perturbations propagate through the backward maps

$$b_K^\mu, \ldots, b_1^\mu.$$

Does the denoising path preserve small discrepancies, or can it amplify them before reaching time zero? What features of μ make the reverse path stable? For multimodal or non-log-concave laws, at which noise scales should denoising be hardest? The chapter develops a statistical perspective on diffusion sampling through these questions, using localization, curvature, and multi-scale complexity to study the geometry of the diffuse-then-denoise procedure.

1.2 Simulating Observed Structure

A statistical account of generative models prepares a second question. Once an underlying data law can be represented by simulation, how can that representation be used? A sampler gives operational access to a fitted law through repeated draws, allowing Monte Carlo evaluation of quantities defined under it. A conditional sampler gives sharper access. By fixing conditioning information and sampling what remains random, the statistician can examine selected parts of an observational law under controlled conditions.

This creates a limited, synthetic analogue of experimental control. It does not intervene on the world; it intervenes only within a fitted observational law. In an experiment, the statistician controls features of the data-generating process and observes the consequences. In synthetic simulation, the statistician fixes the conditioning information inside a fitted law and samples the remaining variables. The first creates comparisons in the real world; the second creates comparisons in an approximated world. The object of control is different, but the logic is parallel: hold fixed the relevant conditions, vary what remains random, and evaluate the quantity of interest.

Conditional moment restrictions are a natural setting for this viewpoint. Many statis-

tical and econometric targets are defined through statements about averages. Given an observation O and conditioning variables S, a parameter or function may be characterized by

$$\mathbb{E}[\psi(O, \theta_0) \mid S] = 0.$$

Such restrictions are attractive because they express identifying content without requiring a full parametric model for the joint distribution. The difficulty is that the restriction must hold across the support of S, so estimation requires control of conditional expectations indexed by the conditioning variables. A sufficiently accurate conditional sampler suggests a direct route: condition on $S = s$, simulate from the fitted conditional law, and evaluate the implied moment.

Instrumental variables provide a canonical example. Suppose we observe

$$(Y_i, X_i, Z_i, W_i), \qquad i = 1, \ldots, n,$$

where Y_i is an outcome, X_i is an endogenous regressor, Z_i is an instrument, and W_i denotes observed covariates. A structural function g_0 is characterized by

$$Y = g_0(X, W) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid Z, W] = 0,$$

or equivalently,

$$\mathbb{E}\{Y - g_0(X, W) \mid Z, W\} = 0.$$

Direct prediction of Y from (X, W) does not target g_0 when X is endogenous. This difficulty has motivated a large literature on IV estimation in econometrics and statistics. Popular methods use sieve or kernel representations and replace the conditional restriction by finite collections of unconditional moments. Our viewpoint is different: we ask whether the conditional restriction can be made computationally accessible by learning the relevant conditional law and simulating from it, for instance through a generative model.

Chapter 5 develops this approach. It studies how simulated draws can be used to evaluate the residual moments defining g_0, thereby making flexible nonlinear specifications operational. This perspective raises several natural questions. How should one formulate a simulation-based criterion that targets the appropriate restriction? How should one couple Monte Carlo draws with estimation of g_0? What notion of first-stage accuracy matters when the first stage is a learned conditional distribution?

1.3 Stabilizing Observed Relations

Generative models provide operational approximations to an observed data law, and simulation makes structures within that law accessible. This is powerful, but necessarily local: it is local to the law that has been learned, and therefore to the observational environment that produced the data. Prediction under distribution shift poses a different question. It asks which features of the source environment remain informative when the environment changes. A relation learned in one population, institution, time period, or deployment setting may be useful where it was observed but fragile elsewhere. The statistical object is therefore no longer the source law itself, but the component of its structure that transfers. The basic domain adaptation problem is as follows. The learner observes labeled source data

$$(X_i^S, Y_i^S) \sim P_S(X, Y),$$

and unlabeled target covariates

$$X_i^T \sim P_T(X).$$

The goal is to construct a predictor with small target risk

$$R_T(f) = \mathbb{E}_T[(Y - f(X))^2],$$

although target labels are unavailable. If the conditional relation between X and Y were stable, source labels could be used to guide target prediction. Observational data make this premise delicate. An environmental shift may reflect changes in unobserved factors that affect both covariates and outcomes. Then the difficulty is not only that $P_S(X) \neq P_T(X)$. The predictive concept itself may change.

Chapter 7 studies this possibility through a new structural model for distribution shift (Dharmakeerthi et al. [2026]). Let Z denote latent exogenous variation that is stable across environments, and let E denote latent environment-specific variation. The covariates and response satisfy

$$X = \Theta Z + \Delta E + W,$$

and

$$Y = \langle \beta^\star, X \rangle + \langle \gamma, E \rangle + U.$$

A shift in the distribution of E changes the observed covariate distribution. Since E also enters the outcome equation, it can also change the best prediction rule. Thus a predictor learned perfectly from the source environment may still be inappropriate for the target environment.

The model raises a natural question. If some latent variation is stable and some is environment-specific, can the observed source labels and target covariates reveal which predictive structure should be trusted? The answer cannot come from source prediction alone, since source accuracy may exploit unstable directions. It also cannot come from distributional stability alone, since stable directions may carry little signal. The chapter asks how this tradeoff should be formalized, when stability of a representation is informative for prediction, and whether a practical learning criterion can recover useful structure under concept shift.

1.4 Organization

The remainder of the thesis is organized as follows.

Chapter 2, *Denoising in Classical Statistics* looks back at the denoising problem as it has appeared in the classical statistics literature. By understanding some fundamental results and perspectives, we ground diffusion models in a foundational statistics framework.

Chapter 3, *Representing Observed Environments: Diffusion Sampling, Denoising, and Geometry*, studies diffusion-based generative sampling. The chapter analyzes the oracle backward path associated with an observational law and identifies geometric quantities governing the stability of the diffuse-then-denoise procedure.

Chapter 4, *Distributional Denoising as Local Transport*, reconsiders the denoising step that lies at the heart of diffuse-then-denoise sampling schemes. The chapter asks what form of denoising is appropriate when the object of interest is not a point estimate, but a recovered distribution. The chapter is prospective and is intended to invite theoretical work on generative models from a measure transport lens.

Chapter 5, *Simulating Observed Structure: Moment Restrictions and Generative Instrumental Variables*, studies nonlinear IV estimation with learned conditional samplers. The chapter formulates a simulation-based conditional moment criterion, develops a Monte Carlo correction for optimization, and relates the feasible criterion based on the learned conditional law to the oracle criterion based on the true conditional law.

Chapter 6, *Generative Bootstrap and Local Environments*, is more retrospective in spirit. It considers generative models as simulation tools within statistical frameworks, and invites further theoretical and applied work on how such models can be used responsibly and effectively.

Chapter 7, *Stabilizing Observed Relations: Confounding, Concept Shift, and Invariant Prediction*, studies prediction under distribution shift with unobserved confounding. The chapter introduces a structural model in which the best source predictor may differ from the best target predictor, and develops a representation-learning approach that trades off source

predictability against stability across environments.

CHAPTER 2

DENOISING IN CLASSICAL STATISTICS

The first substantive chapter studies diffusion-based generative modeling as a way of representing an observed data distribution. Before turning to that analysis, it is useful to return to a more classical problem. Diffusion models are often framed through the language of stochastic differential equations, and measure-valued dynamics. These perspectives are essential, and the next chapter develops them formally. Yet the local operation underlying the procedure is much older and simpler. It is denoising: one observes a corrupted version of a signal and uses information in the corrupted distribution to move backward toward the signal.

The additive-noise model

$$Y = X + \sigma Z$$

is among the simplest models in statistics. It has been studied in many forms: as a measurement-error model, as a deconvolution problem, and as a decision-making problem. It also provides the local statistical template for diffusion models. Each "forward" step of a diffuse–then–denoise sampling scheme creates a noisy observation of a less noisy state, and the "backward" procedure asks how that corruption should be undone.

$$\text{(Diffuse)} \quad X \to Y.$$

$$\text{(Denoise)} \quad Y \to X.$$

This chapter identifies the statistical lineage behind that local operation. We begin with the classical problem of recovering a signal distribution from additive noise, then turn to pointwise recovery and empirical Bayes, where Tweedie's formula reveals the score of the noisy marginal distribution as the object governing Gaussian denoising. Score matching then provides a way to estimate this object directly from data. Diffusion models combine

these ideas in a new role: they learn noisy marginal scores across a path of noise levels and compose the resulting local denoisers into a sampler.

2.1 Distributional recovery: deconvolution

The local denoising problem in a diffusion model is closely related to a classical problem of distributional recovery. At a single noise level, one observes samples from a corrupted distribution and seeks information about the less noisy distribution from which they arose. In the additive setting, this is the deconvolution problem.

For simplicity, consider the one-dimensional model

$$X_1, \dots, X_n \overset{\text{iid}}{\sim} f_X, \qquad Y_i = X_i + Z_i,$$

where the measurement errors Z_i are independent of the X_i and have known density f_Z. The statistician observes $Y_1, \dots, Y_n$, not $X_1, \dots, X_n$. The observed density is therefore

$$f_Y(y) = \int_{\mathbb{R}} f_Z(y - x) f_X(x)\, dx = (f_Z * f_X)(y). \tag{2.1.1}$$

The deconvolution problem is to recover the latent density f_X from observations drawn from the convolved density f_Y. This is one of the classical statistical formulations of distributional recovery under noise.

Our treatment in this introductory chapter is deliberately limited. The goal is not to survey the deconvolution literature in full, but to identify the nature of the statistical problem and its difficulty. We follow standard accounts of statistical deconvolution, including Fan [1991], Meister [2009], and the exposition of Panaretos [2011].

The convolution structure in (2.1.1) suggests a direct inversion strategy. Let φ_X, φ_Y, and φ_Z denote the characteristic functions of X, Y, and Z, respectively. Then

$$\varphi_Y(u) = \varphi_X(u)\varphi_Z(u).$$

If $\varphi_Z(u) \neq 0$, the formal inverse is

$$\varphi_X(u) = \frac{\varphi_Y(u)}{\varphi_Z(u)}, \qquad f_X(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-iux} \frac{\varphi_Y(u)}{\varphi_Z(u)}\, du. \tag{2.1.2}$$

This display gives the problem its apparent simplicity. Estimate φ_Y from the noisy observations, divide by the known noise characteristic function, and invert the Fourier transform.

The difficulty is that this inverse operation is unstable. Convolution smooths. Its inverse must recover the high-frequency information that smoothing suppresses. If $\widehat{\varphi}_Y(u)$ is an empirical estimate of $\varphi_Y(u)$, the deconvolution error contains misbehaved terms. Observe that, by the Plancherel identity, we have,

$$\int \left| f_X(x) - \hat{f}_X(x) \right|^2 dx = \int \left| \phi_X(u) - \frac{\hat{\phi}_Y(u)}{\phi_Z(u)} \right|^2 du = \int \left| \frac{\phi_Y(u) - \hat{\phi}_Y(u)}{\phi_Z(u)} \right|^2 du$$

At frequencies where $|\varphi_Z(u)|$ is small, even modest estimation error in $\widehat{\varphi}_Y(u)$ can be greatly amplified. Gaussian errors are especially severe: their characteristic functions decay exponentially, so high-frequency recovery becomes strongly ill-posed. This instability is the central statistical obstruction in nonparametric deconvolution. A large statistical literature studies how to regularize this inverse problem and how the resulting rates depend on the smoothness of the signal density and the decay of the noise characteristic function [Carroll et al., 2006, Meister, 2009, Delaigle and Gijbels, 2004].

Deconvolution is a classical theory of distributional recovery. It asks for the latent distribution and gives precise statistical meaning to the difficulty of recovering that distribution from a corrupted marginal. But its output is a representation of the density or distribution of X. For diffusion models, this is not quite the operational object required. A sampler needs a mechanism for moving samples. It needs a map, transition, or transformation that takes a noisy draw and moves it toward a less noisy distribution. Deconvolution estimates what the latent distribution is; diffusion models ask how to move toward that distribution through a sequence of denoising operations.

This distinction motivates a second classical perspective. Instead of trying first to recover the entire density f_X, one may ask for the best action to take after observing a noisy value $Y = y$. That is the empirical-Bayes point of view. It leads to posterior denoising rules; and in the Gaussian additive model, those rules are governed by the score of the noisy marginal distribution.

2.2 Pointwise recovery: empirical Bayes and Tweedie's formula

A second classical interpretation of the same Gaussian model is pointwise recovery. The target is no longer the full distribution P_X, but the latent realization paired with a noisy observation. Given

$$Y_i = X_i + \sigma Z_i, \qquad Z_i \sim \mathcal{N}(0, I_d),$$

the statistician seeks a map T such that $T(Y_i)$ is close to X_i. This is closer to the local denoising problem that appears inside diffusion models. A denoising map should act on noisy samples and move them toward less noisy ones.

In the decision-theoretic formulation, the statistician chooses a measurable map T and evaluates it through a loss. Under squared-error loss, the risk is

$$\mathcal{R}(T) = \mathbb{E}\|X - T(Y)\|^2.$$

The Bayes rule is the posterior mean,

$$T^\star(y) = \mathbb{E}[X \mid Y = y]. \tag{2.2.1}$$

This fact is elementary, but it identifies the statistical object. Under squared-error loss, denoising means estimating the conditional mean of the latent signal given the noisy observation.

At first sight, (2.2.1) requires knowledge of the prior distribution of X. If P_X were known,

the posterior mean could be computed by Bayes' rule. The empirical-Bayes perspective asks how much of this Bayes behavior can be recovered from repeated observations alone. Robbins' empirical-Bayes program showed that prior-driven quantities can often be estimated from the ensemble of observations: the marginal distribution of the noisy sample carries information about the population from which the latent signals arise, and that marginal distribution can determine the Bayes action [Robbins, 1956, 1964].

In the Gaussian additive model, this route is especially elegant. Let q denote the marginal density of Y. Tweedie's formula gives

$$\mathbb{E}[X \mid Y = y] = y + \sigma^2 \nabla \log q(y). \tag{2.2.2}$$

The Bayes rule appears to depend on the unknown distribution of X, but in the Gaussian denoising problem the required correction is the score of the noisy distribution,

$$\nabla \log q(y).$$

This identity is the bridge from empirical Bayes to score-based generative modeling. The prior need not be recovered explicitly in order to construct the pointwise Bayes rule; it is enough, formally, to know the score of the corrupted marginal distribution. Diffusion models will use the same object, not once, but across a path of noise levels.

Empirical Bayes and shrinkage. While tangential, it is worthwhile at this point to mention a remarkable connection between the empirical-Bayes perspective and shrinkage. The James–Stein phenomenon shows that, in multivariate Gaussian mean estimation, the usual unbiased estimator can be dominated under quadratic loss by a rule that "shrinks" the observation toward a central point [James and Stein, 1961]. Efron and Morris interpreted this phenomenon through empirical Bayes: the amount of shrinkage is learned from the ensemble rather than fixed in advance [Efron and Morris, 1973].

Their interpretation is closely aligned with Tweedie's formula. In the Gaussian sequence problem, the James–Stein estimator can be written as a score-corrected rule

$$T^{\mathrm{JS}}(y) = y + \sigma^2 \widehat{s}(y),$$

where the empirical score estimate takes the form $\widehat{s}(y) = -\frac{d-2}{\|y\|^2} y$ and thus,

$$T^{\mathrm{JS}}(y) = \left(1 - \frac{(d-2)\sigma^2}{\|y\|^2}\right) y.$$

In this sense, James–Stein shrinkage may be viewed as an empirical-Bayes score correction: the observation is adjusted according to information carried by the marginal distribution of the observed vector.

This is the lesson that will reappear in diffusion models. Score-based methods that are utilized in modern samplers are continuous with a classical statistical idea. The score of a noisy distribution governs how observations should be moved when the goal is denoising.

2.3 Score estimation and score matching

Tweedie's formula reduces Gaussian empirical-Bayes denoising to score estimation. The remaining question is how the score is to be learned from data. One route is to estimate the density q and then differentiate $\log q$. In flexible or high-dimensional models, however, density estimation is itself a difficult task, and normalized likelihoods may be unavailable or inconvenient.

Hyvärinen's score matching criterion was introduced precisely for this purpose [Hyvärinen, 2005]. In its simplest population form, for a vector field $s : \mathbb{R}^d \to \mathbb{R}^d$ and $Y \sim q$, define

$$\mathcal{J}(s) = \mathbb{E}_q \left[\frac{1}{2}\|s(Y)\|^2 + \nabla \cdot s(Y)\right]. \tag{2.3.1}$$

Under suitable boundary conditions, integration by parts shows that the minimizer is

$$s^\star(y) = \nabla \log q(y).$$

The normalizing constant of q does not appear, and the criterion targets the score directly. Therefore, the vector field needed for denoising can be learned without explicitly recovering either the prior distribution of X or a normalized density for the noisy marginal.

The same derivative structure appears in Stein's unbiased risk estimate. Write a Gaussian denoising rule as

$$T(y) = y + h(y).$$

Then, formally, Stein's identity gives

$$\mathbb{E}\|X - T(Y)\|^2 = \sigma^2 d + \mathbb{E}\|h(Y)\|^2 + 2\sigma^2 \mathbb{E}[\nabla \cdot h(Y)]. \tag{2.3.2}$$

A second integration by parts rewrites the divergence term as

$$\mathbb{E}[\nabla \cdot h(Y)] = -\mathbb{E}[h(Y) \cdot \nabla \log q(Y)]\,.$$

Thus risk, denoising, and the marginal score are linked by the same integration-by-parts identities. Empirical Bayes explains why the score determines the pointwise Gaussian denoising rule. Stein identities explain why this score enters risk calculations. Score matching explains how the score can be estimated directly.

In the next section we will explain how diffusion models can be interpreted as an iterated denoising procedure, where a score-based denoising map is required across a range of noise levels along a corruption path. Despite the introduced complexity, the local statistical object will remain the same: the score of a noisy marginal distribution. What changes is its role.

2.4 From denoising to diffusion sampling

Diffusion models change the role of denoising. In empirical Bayes, the denoising rule is the final action: after observing Y, one reports $T(Y)$. In a diffusion model, the denoising rule is one local operation inside a generative chain.

At a high level, a diffusion model couples two procedures. The first is a forward noising process, which gradually transforms the data distribution into a tractable reference distribution, typically a Gaussian. If $X_0 \sim \mu$ denotes an observation from the data distribution, the forward process constructs a path of distributions

$$\mu = \mu_0 \longrightarrow \mu_{t_1} \longrightarrow \cdots \longrightarrow \mu_{t_K} \approx \nu,$$

where ν is easy to sample from. The second procedure is the reverse operation. Sampling begins with a draw from, or near, ν, and then applies learned denoising steps intended to move the sample through progressively less noisy distributions:

$$\nu \approx \mu_{t_K} \longrightarrow \mu_{t_{K-1}} \longrightarrow \cdots \longrightarrow \mu_{t_1} \longrightarrow \mu_0 = \mu.$$

In this sense, generation is organized as a diffuse–then–denoise procedure. The forward path makes the original distribution simple; the reverse path attempts to recover the original distribution by repeatedly undoing small corruptions.

Diffusion models are often described with the language of stochastic differential equations, or probability flow ordinary differential equations. The next chapter develops a measure-transport perspective on this construction. For the present purpose, however, it is enough to examine a single local step. At that scale, the classical Gaussian denoising calculation reappears.

Consider the discretized Ornstein–Uhlenbeck step frequently employed in "forward" dif-

fusion,

$$X_{t+1} = (1-\eta)X_t + \sqrt{2\eta}\, Z, \qquad Z \sim N(0, I_d).$$

Writing

$$Y = X_{t+1}, \qquad \widetilde{X} = (1-\eta)X_t,$$

we obtain the additive Gaussian model

$$Y = \widetilde{X} + \sqrt{2\eta}\, Z.$$

Thus a single reverse step asks for information about a less noisy state $\widetilde{X}$ from a noisy observation Y. If q_{t+1} denotes the density of Y, Tweedie's formula gives

$$\mathbb{E}[\widetilde{X} \mid Y = y] = y + 2\eta \nabla \log q_{t+1}(y).$$

The one-step backward mean is therefore governed by the score of the noisy distribution. We will soon see that the "backward" denoising procedure employed at sampling time involves a repeated application of score-based rules at differing noise scales.

A denoising diffusion model may therefore be viewed as a multiscale extension of the classical denoising problem. The single noisy marginal density q is replaced by a path of noisy marginals $\{q_t\}$, and the single score $\nabla \log q$ is replaced by a time-indexed score field

$$s_t(x) = \nabla \log q_t(x).$$

The denoiser is no longer the final statistical action. It becomes a local operation inside a composed sampler. The next chapter studies this composition directly.

CHAPTER 3

REPRESENTING OBSERVED ENVIRONMENTS: DIFFUSION SAMPLING, DENOISING, AND GEOMETRY

3.1 Introduction

Empirically, diffusion models exhibit compelling performance as probabilistic generative models for complex, multi-dimensional probability measures [Ho et al., 2020, Sohl-Dickstein et al., 2015, Song and Ermon, 2019, Song et al., 2021, Karras et al., 2022]. They are often employed when traditional sampling methods suffer, such as when the probability measure is multi-modal and supported on an unknown manifold that is hard to mathematize, such as the probability distribution of (pixels of) images. Despite their impressive performance in practice, several fundamental theoretical questions regarding denoising quality remain unanswered [Block et al., 2020, Chen et al., 2022, Lee et al., 2022].

In a nutshell, diffusion models aim to revert a Langevin diffusion chain, moving from a log-concave equilibrium measure ν, say an isotropic Gaussian, back toward a complex, possibly non-log-concave initial measure μ. A Langevin diffusion is a forward chain in the space of probability measures, implemented iteratively with a stepsize η as in (3.1.1), where ν is the equilibrium measure and $\mathbf{f}_k^\mu$ denotes the forward transition map[1] at step k.

$$\text{Forward Diffusion} \quad \mu =: \mu_0 \overset{\mathbf{f}_1^\mu}{\to} \mu_\eta \to \cdots \overset{\mathbf{f}_k^\mu}{\to} \mu_{k\eta} \to \cdots \overset{\mathbf{f}_K^\mu}{\to} \mu_{K\eta} \quad \overset{K\to\infty}{\rightsquigarrow} \nu \,, \tag{3.1.1}$$

$$\text{Time Reversal} \quad \mu =: \mu_0 \underset{\mathbf{b}_1^\mu}{\leftarrow} \mu_\eta \leftarrow \cdots \underset{\mathbf{b}_k^\mu}{\leftarrow} \mu_{k\eta} \leftarrow \cdots \underset{\mathbf{b}_K^\mu}{\leftarrow} \mu_{K\eta} \quad \boxed{\underset{?}{\leftarrow} \cdots \underset{\mathbf{b}^\nu}{\leftarrow} \nu} \,. \tag{3.1.2}$$

This forward chain is Markovian and thus time-reversible as in (3.1.2), where $\mathbf{b}_k^\mu$ denotes the backward transition map[2] for the μ-chain at step k. For sampling, diffusion models propose starting the time reversal process at $K \to \infty$, namely the equilibrium measure ν, and aim

1. The rigorous definition will follow in Section 3.3.

2. Again, the rigorous definition will follow in Section 3.3.

to reverse it back to the initial measure μ. However, this is problematic both theoretically (in terms of probability) and conceptually (in terms of optimization). Theoretically, the time reversal of a Markov Chain starting from an equilibrium (invariant measure) will get stuck [Norris, 1998]. The backward chain $\nu \leftarrow_{\mathbf{b}^\nu} \nu$ will stay as the invariant measure and never reach $\mu_{K\eta}$. Conceptually, hoping to trace back the initial condition μ starting from ν is infeasible. If two forward chains with initials $\mu \neq \mu'$ both end up at ν, the time-reversal starting from ν (and solely using the future information) cannot recover the past and distinguish these two chains.

Therefore, framing the diffusion model as a time-reversal Markov chain—one that recovers the past from the future—does not fully resolve the underlying conceptual issues. In contrast, we propose studying diffusion models as a form of *sensitivity analysis*. It is clear that starting from $\mu_{K\eta}$ and traversing back following transitions $\mathbf{b}^\mu_K \cdots \mathbf{b}^\mu_k \cdots \mathbf{b}^\mu_1$ will identify μ. But what will happen if we start from a perturbed version $\nu =: \bar{\nu}_{K\eta} \neq \mu_{K\eta}$ and follow the same traversing path $\mathbf{b}^\mu_K \cdots \mathbf{b}^\mu_k \cdots \mathbf{b}^\mu_1$ as illustrated in (3.1.4)?

$$\text{Time Reversal} \quad \mu =: \mu_0 \underset{\mathbf{b}^\mu_1}{\leftarrow} \mu_\eta \leftarrow \cdots \underset{\mathbf{b}^\mu_k}{\leftarrow} \mu_{k\eta} \leftarrow \cdots \underset{\mathbf{b}^\mu_K}{\leftarrow} \mu_{K\eta} \quad \boxed{\leftarrow \cdots \leftarrow \nu \;\; ? \;\; \mathbf{b}^\nu} \tag{3.1.3}$$

$$\text{Backward Denoising} \quad \mu \overset{?}{\approx} \bar{\nu}_0 \underset{\mathbf{b}^\mu_1}{\leftarrow} \bar{\nu}_\eta \leftarrow \cdots \underset{\mathbf{b}^\mu_k}{\leftarrow} \bar{\nu}_{k\eta} \leftarrow \cdots \boxed{\underset{\mathbf{b}^\mu_K}{\leftarrow} \bar{\nu}_{K\eta} := \nu}, \tag{3.1.4}$$

The traversing path carries the past information, namely the signature of the initial measure μ, but starts with an easy-to-sample ν as a surrogate, replacing $\mu_{K\eta}$. This mismatch renders the backward denoising process (3.1.4) *non-Markovian*, thereby making it plausible to recover the past from the future.

Sensitivity Analysis: Score, Curvature, and Localization. How can the traversing operator $\mathbf{b}^\mu_k$ be estimated? We show in Proposition 3.2.3 that the optimal backward operator $\mathbf{b}^\mu_k$—in terms of transportation cost—depends on the score function $\nabla \log p_{\mu_{k\eta}}$ where $p_{\mu_{k\eta}}$ is the probability density function of $p_{\mu_{k\eta}}$. A key observation in diffusion models is that score

estimation can be cast as a supervised prediction problem [Saremi and Hyvärinen, 2019] using the forward diffusion chain, where the goal is to predict the previous location given the current one, in terms of conditional expectation, as seen in Proposition 3.2.4.

The fundamental question is whether the perturbation $\bar{\nu}_{K\eta} \approx \mu_{K\eta}$ will get amplified along the backward denoising chain. Will $\bar{\nu}_{k\eta}$ as in (3.1.3)-(3.1.4) stay close to $\mu_{k\eta}$, for $k = K, \cdots, 0$? This chapter provides a precise study of diffusion models from the viewpoint of denoising quality. We unveil in Proposition 3.2.5 that the curvature function $\nabla^2 \log p_{\mu_{k\eta}}$ controls the key aspects of this sensitivity analysis. In other words, the curvature function governs the score function's denoising capability, which we refer to as localization. Localization quantifies the uncertainty of the previous location given the current, in terms of conditional covariance.

Most of the current theoretical literature [Lee et al., 2023, Chen et al., 2022, 2023] treats this curvature as a nuisance, for example, by focusing on "non-expansion" metrics d, and leveraging a form of data-processing inequality,

$$d(\mu_{k-1}, \bar{\nu}_{k-1})/d(\mu_k, \bar{\nu}_k) \leq 1, \text{ for } d \in \{d_{\mathrm{TV}}, d_{\mathrm{KL}}\} .$$

Consequently, $d(\mu_0, \bar{\nu}_0) \leq d(\mu_K, \nu)$, and $d(\mu_0, \bar{\nu}_0)$ can be determined by the contraction of the forward diffusion chain alone. However, even for simple Gaussians, the backward denoising could either be (i) an expansion or (ii) a contraction much faster than the forward diffusion, under the Wasserstein-2 metric, W. Consider the forward diffusion (3.1.5) with initialization $\mu \sim \mathcal{N}(0, s^2)$ and temperature $\beta = 1$. Recall, the forward diffusion contracts in W at a rate $1 - \eta$. Our Corollary 3.3.5 implies that the one-step backward denoising at

effective time $t = k\eta$ with $\mu_k \sim \mathcal{N}(e^{-k\eta}, e^{-2k\eta}s^2 + 1 - e^{-2k\eta})$ satisfies the equality

$$\frac{W(\mu_{k-1}, \bar{\nu}_{k-1})}{W(\mu_k, \bar{\nu}_k)} = 1 + \eta \frac{s^2-1}{e^{k\eta}+s^2-1} \ , \quad \begin{cases} (i) > 1 & \text{if } s^2 > 1, \\ (ii) < 1 - \eta & \text{if } s^2 < \frac{1}{2}, \text{ and} \\ & k < \frac{1}{2\eta}\log\big(2(1-s^2)\big). \end{cases}$$

In either case, treating the backward denoising as simply non-expansive under certain metrics is losing significant information about the diffuse-then-denoise process.

In contrast, we study the process under the Wasserstein-2 metric and provide a fine-grained analysis of how a complexity measure based on the curvature, $\nabla^2 \log p_{\mu_{k\eta}}$, completely governs the expansion or contraction of the backward denoising step.

Diffuse-then-Denoise Process: Offset Contraction/Expansion. The sensitivity analysis is equivalent to studying the effectiveness of a diffuse-then-denoise process. To simplify the exposition, we consider a one-step version here. Given any two measures, $\mu_0 \neq \nu_0$, run one step of forward diffusion as in (3.1.1) with $K = 1$, with $\mathbf{Z}$ isotropic Gaussian and $\beta \in \mathbb{R}_+$, the temperature,

$$\mathbf{X}_\eta = (1-\eta)\mathbf{X}_0 + \sqrt{2\beta^{-1}\eta}\mathbf{Z}, \ \mathbf{X}_0 \sim \mu_0, \ \text{ and } \ \mathbf{Y}_\eta = (1-\eta)\mathbf{Y}_0 + \sqrt{2\beta^{-1}\eta}\mathbf{Z}, \ \mathbf{Y}_0 \sim \nu_0 \ .$$

Denote the measure associated with $\mathbf{Y}_\eta \sim \bar{\nu}_\eta$, then run one step of backward denoising with the optimal $\mathbf{b}^\mu$ (as in (3.1.4) with $K = 1$) to obtain $\bar{\nu}_0 \leftarrow_{\mathbf{b}^\mu} \bar{\nu}_\eta$. A natural question is whether the cumulative effect of this diffuse-then-denoise process is a contraction. Namely, is $W(\mu_0, \bar{\nu}_0)$ smaller than $W(\mu_0, \nu_0)$?

At first sight, it may seem unnatural that the diffuse-then-denoise process will yield an improvement. How can adding noise and then denoising be better? Consider setting $\beta^{-1} = 0$, and let $\mu_0 = \delta_x$ and $\nu_0 = \delta_y$ be two Dirac measures supported at $x \neq y$. One can

verify that

$$\text{Forward } W(\mu_0,\nu_0) = \|x-y\| \ , \ \text{Backward } W(\mu_\eta,\bar{\nu}_\eta) = (1-\eta)\|x-y\| \ ,$$
$$\text{Forward-then-Backward } W(\mu_0,\bar{\nu}_0) = (1-\eta)^{-1}W(\mu_\eta,\bar{\nu}_\eta) = \|x-y\| = W(\mu_0,\nu_0) \ .$$

Here, the backward expansion offsets the forward contraction, resulting in no net effect for the forward-then-backward process.

Curiously, as we shall show in Theorems 3.3.1, 3.3.3 and 3.3.6, roughly speaking, when $\beta^{-1} \neq 0$ and $\nabla^2 \log p_\mu$ has a certain curvature complexity quantified by $\zeta \in \mathbb{R}$, there is a net effect of the diffuse-then-denoise process

$$\begin{aligned}
&\text{Forward Diffusion:} && \frac{W(\mu_\eta,\bar{\nu}_\eta)}{W(\mu_0,\nu_0)} \leq 1-\eta+O(\eta^2) \ , \\
&\text{Backward Denoising:} && \frac{W(\mu_0,\bar{\nu}_0)}{W(\mu_\eta,\bar{\nu}_\eta)} \leq 1+\eta-\eta\beta^{-1}\zeta+O(\eta^2) \ , \\
&\text{Diffuse-then-Denoise:} && \frac{W(\mu_0,\bar{\nu}_0)}{W(\mu_0,\nu_0)} \leq 1-\eta\beta^{-1}\zeta+O(\eta^2) \ .
\end{aligned}$$

This contrasts sharply with the $\beta^{-1} = 0$ case: curiously, adding non-trivial noise then denoising could be beneficial and result in an effective net contraction $\beta^{-1}\zeta$, provided the curvature $\zeta > 0$. We emphasize that these inequalities become equalities for simple log-concave μ_0's, thus establishing the sharpness of our characterization, see Corollary 3.3.5. In summary, the success or failure of the diffuse-then-denoise process solely depends on the curvature/localization function defined in Proposition 3.2.5.

Chaining the argument, we show in Corollary 3.3.7 that, rather than log-concavity, it is a multi-scale complexity along all time scales that controls the net effect of the diffuse-then-denoise process (3.1.4) with K steps

$$\frac{W(\mu_0,\bar{\nu}_0)}{W(\mu_0,\nu_0)} \leq \exp\left(-\beta^{-1}\eta\sum_{k=1}^{K}\zeta_{k\eta}+O(\eta^2)\right) \ ,$$

where $\zeta_{k\eta}$ is some curvature/localization complexity of $\mu_{k\eta}$ at time $t = k\eta$. For a general initial measure μ, the diffused version μ_t could be non-log-concave, depending on the time scale t. In Section 3.4, we introduce a notion of multi-scale complexity that describes the localization difficulty for generic μ and is new to the literature. The multi-scale corresponds to the effective signal-to-noise ratio for the denoising step at each time scale t.

Multi-Scale Complexity: Beyond Log-Concavity. With any initial measure μ, the Ornstein-Uhlenbeck diffusion process at different time scales t is closely tied to a multi-scale smoothing, at different signal-to-noise ratio $\mathsf{r} = \mathsf{r}(t)$ (defined in (3.4.2)),

$$\mathbf{Y}_{\mathsf{r}} := \mathsf{r}\mathbf{X} + \mathbf{Z} \ , \ (\mathbf{X}, \mathbf{Z}) \sim \mu \otimes \mathcal{N}(0, 1) \ .$$

We introduce in Section 3.4 the following multi-scale complexity, defined based on the localization function $L_{\mathsf{r}}(y) := \| \operatorname{Cov}[\mathsf{r}\mathbf{X}|\mathbf{Y}_{\mathsf{r}} = y]\|_{\mathrm{op}}$,

$$h_\mu(\delta, \mathsf{r}) = \int_{1-\delta}^{\infty} \mathbb{P}\big(L_{\mathsf{r}}(\mathbf{Y}_{\mathsf{r}}) > u\big) \mathrm{d}u \ , \ \delta \in (0, 1] \ .$$

Here $\mathbb{P}\big(L_{\mathsf{r}}(\mathbf{Y}_{\mathsf{r}}) > u\big)$ is the tail probability of random variable $\| \operatorname{Cov}[\mathsf{r}\mathbf{X}|\mathbf{Y}_{\mathsf{r}}]\|_{\mathrm{op}}$, and therefore $h_\mu(\delta, \mathsf{r})$ controls its integrated tail. We show in Theorem 3.4.6 that the growth of this integrated tail function $\delta \mapsto h_\mu(\delta, \mathsf{r})$ governs the effectiveness of the diffuse-then-denoise process at any time t, with the corresponding SNR scale $\mathsf{r}(t)$. Our characterization holds for any generic μ that extends far beyond log-concavity.

This multi-scale complexity at every SNR scale $\mathsf{r} \geq 0$, conceptually, is quantifying the curvature $\nabla^2 \log p_{\mathbf{Y}_{\mathsf{r}}}(y)$ in a certain average sense weighted by $p_{\mathbf{Y}_{\mathsf{r}}}(y)$. This notion, rather than the worst-case curvature, governs the localization accuracy of the backward denoising chain at every scale. Our analysis is fine-grained in two senses. First, for any $\mathbf{X} \sim \mu$, across different scales of r, the non-log-concavity of $p_{\mathbf{Y}_{\mathsf{r}}}$ changes in a complex, multi-resolution way. Non-log-concavity across all scales of r collectively determines the effect of the backward

denoising chain. Second, unlike in the worst-case analysis, where the worst point y with the largest positive curvature $\nabla^2 \log p_{\mathbf{Y}_\mathsf{r}}(y)$ dictates the analysis, we leverage the following observation. If a point y with a large positive curvature $\nabla^2 \log p_{\mathbf{Y}_\mathsf{r}}(y)$ is unlikely to occur with $p_{\mathbf{Y}_\mathsf{r}}(y)$ small, the overall non-log-concavity is still benign. This is true in many examples; see Section 3.5.

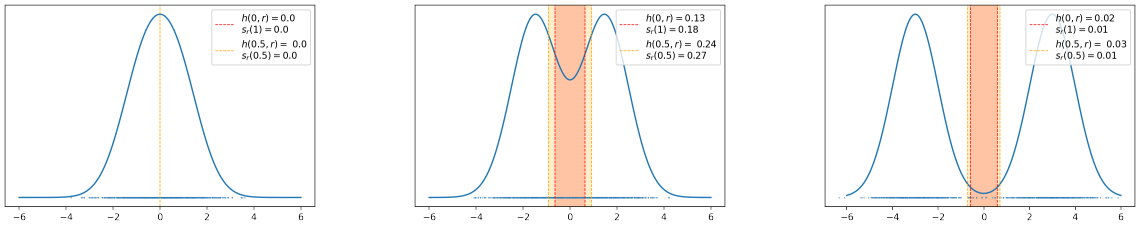
This multi-scale complexity may appear mysterious; we present a concrete, non-log-concave example to clarify intuitions. Consider the simplest one-dimensional non-log-concave measure $\mu = \frac{1}{2}\delta_{-1} + \frac{1}{2}\delta_{+1}$.

- Localization and Curvature: $L_\mathsf{r}(y) \leq 1-\delta$ if and only if $\nabla^2 \log p_{\mathbf{Y}_\mathsf{r}}(y) \preceq -\delta \cdot I_d$, that is, $p_{\mathbf{Y}_\mathsf{r}}(y)$ strongly log-concave at y. These y's are good locations with strong curvature: accurate denoising and localization from y is easy. These are locations where diffuse-then-denoise is beneficial. See locations outside the shaded areas in Figure 3.1.

- Survival Function: provided we have samples $\mathbf{Y}_\mathsf{r}$, the survival function $s_\mathsf{r}(1-\delta) := \mathbb{P}\big(L_\mathsf{r}(\mathbf{Y}_\mathsf{r}) > 1-\delta\big)$ tells us the probability of bad locations with possible non-log-concavity where the backward denoising is hard. It quantifies the mass that may induce a large expansion in the diffuse-then-denoise process. See shaded areas in Figure 3.1.

- Integrated Tail: slow growth in the integrated tail function $\delta \to h_\mu(\delta, \mathsf{r})$ implies that one can take an effectively large δ such that the bad locations with positive curvature have a negligible expansion effect, and good locations with strong negative curvature induce a contraction effect, offsetting the expansion. This complexity quantifies an overall notion of curvature.

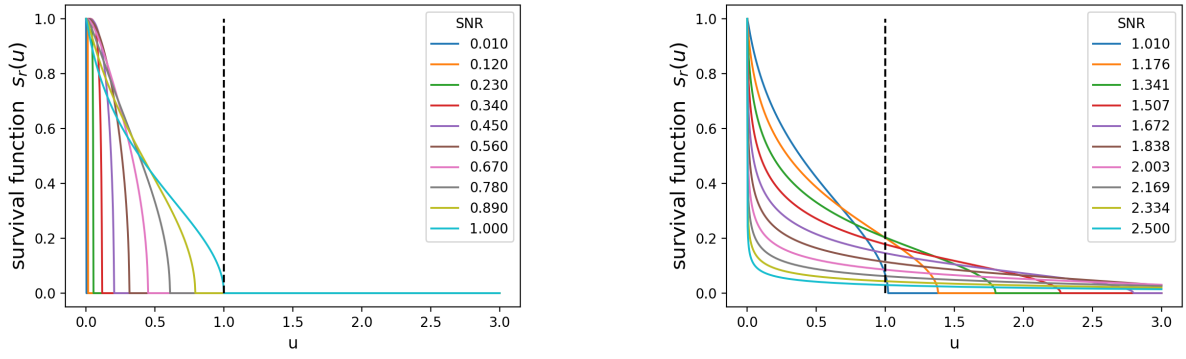
  Figure 3.1 shows: (a) low $\mathsf{r} = 0.71$, $h_\mu(0, \mathsf{r}) = h_\mu(0.5, \mathsf{r}) = 0$, no growth, (b) mid $\mathsf{r} = 1.50$, $h_\mu(0, \mathsf{r}) = 0.13, h_\mu(0.5, \mathsf{r}) = 0.24$, rapid growth of integrated tail as δ increases from 0 to 0.5, and (c) high $\mathsf{r} = 3.00$, $h_\mu(0, \mathsf{r}) = 0.02, h_\mu(0.5, \mathsf{r}) = 0.03$, a very slow growth. Curiously, the complexity is non-monotonic in SNR r: the mid r presents the hardest non-log-concavity for localization.



(a) $\mathsf{r} = 0.71$ *(b)* $\mathsf{r} = 1.50$ *(c)* $\mathsf{r} = 3.00$

Figure 3.1: We plot the density $p_{\mathbf{Y}_\mathsf{r}}(\cdot)$, for three SNR r's. Red shaded area corresponds to non-log-concave region with $\nabla^2 \log p_{\mathbf{Y}_\mathsf{r}}(\cdot) > -\delta$ with $\delta = 0$, and Orange shaded area corresponds to $\delta = 0.5$. For each δ, we report the integrated tail $h_\mu(\delta, \mathsf{r})$ and survival function $s_\mathsf{r}(1-\delta)$ for $\delta \in \{0, 0.5\}$. (a) low $\mathsf{r} = 0.71$, $s_\mathsf{r}(1) = 0$, $s_\mathsf{r}(0.5) = 0$; (b) mid $\mathsf{r} = 1.50$, $s_\mathsf{r}(1) = 0.18$, $s_\mathsf{r}(0.5) = 0.27$, non-trivial mass of bad locations; (c) high $\mathsf{r} = 3.00$, $s_\mathsf{r}(1) = 0.01$, $s_\mathsf{r}(0.5) = 0.01$, though bad locations do exist, samples $\mathbf{Y}_\mathsf{r}$ rarely end up there.

How is this multi-scale complexity useful? Here we plot the survival function $s_\mathsf{r}(u)$, indexed by the different SNR r. Curiously, the growth rate of the integrated tail complexity $h_\mu(\delta, \mathsf{r})$ is non-monotonic in SNR r: both low SNR $\mathsf{r} \le 1$ and high $\mathsf{r} \ge 2$ have extremely slow growth in the integrated tail; the hardest non-log-concavity happens when $\mathsf{r} \in (1, 2)$. Conceptually, for any given initial measure μ, the multi-scale complexity tells us precisely at what time scale $\mathsf{r}(t)$ the backward denoising chain suffers the most. See Section 3.5 for more comprehensive examples.



(a) *(b)*

Figure 3.2: (a) $s_\mathsf{r}(u)$ Low SNR. (b) $s_\mathsf{r}(u)$ High SNR.

3.1.1 Preliminaries and Notations

Notation Throughout this chapter, we consider $\mu \in \mathcal{P}_2(X)$, the space of probability measures with a bounded second moment with $X \subseteq \mathbb{R}^d$. $\mathcal{P}_2^r(X) \subset \mathcal{P}_2(X)$ denotes probability measures absolutely continuous to the Lebesgue measure. For a measure $\nu \in \mathcal{P}_2^r(X)$, denote $\nu = p_\nu \cdot \mathcal{L}^d$ the Radon-Nikodym derivative w.r.t Lebesgue measure, where $p_\nu \in C^1(X)$ is the density function. We reserve $\mathcal{G}, \mathcal{F}, \mathcal{E}$ for functionals $\mathcal{P}_2(X) \to \mathbb{R}$, and h, g, f for real-valued functions $X \to \mathbb{R}$. Let $\boldsymbol{\xi} \in C_c^\infty(X; X)$ denote smooth vector fields; $\mathbf{i}$ denotes the identity map, $\mathbf{t}, \mathbf{f}, \mathbf{b}$ denotes the optimal transport maps, in Definition 3.2.2. We use $\mathbf{X}, \mathbf{Y}, \mathbf{Z}, \mathbf{B}$ to denote random vectors. For a matrix M, $\|M\|_{\text{op}}$ denotes the operator norm, $\text{tr}(M)$ denotes the trace of the matrix. For a vector v, $\|v\|$ denotes the Euclidean norm.

The forward Langevin diffusion process is a stochastic differential equation in the form,

$$\mathrm{d}\mathbf{X}_t = -\nabla f(\mathbf{X}_t)\mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}\mathbf{B}_t, \quad \nabla f \in C_c^\infty(X, X) \; . \tag{3.1.5}$$

Here β^{-1} serves as the temperature parameter, which controls the magnitude of the injected noise $\mathrm{d}\mathbf{B}_t$. The probability measure of $\mathbf{X}_t$, denoted as μ_t with density $\rho_t := p_{\mu_t}$, evolves according to the Fokker-Planck partial differential equation

$$\partial_t \rho_t = \nabla \cdot \big(\rho_t(\nabla f + \beta^{-1}\nabla \log \rho_t)\big) \; . \tag{3.1.6}$$

Forward diffusion can also be viewed as a gradient flow in the Wasserstein space. For two measures $\mu, \nu \in \mathcal{P}_2^r(X)$, define the Wasserstein metric as

$$W_2(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \left(\int \|x - y\|^2 \mathrm{d}\pi(x, y) \right)^{1/2} \; .$$

Jordan et al. [1998] showed that the forward Fokker-Planck PDE, $\mu_t \to \mu_{t+\eta}$, can be viewed as the steepest descent with respect to the Wasserstein metric in the infinitesimal limit, as

$\eta \to 0$

$$\mu_{t+\eta} := \operatorname*{arg\,min}_{\nu \in \mathcal{P}_2^r(X)} \frac{1}{2\eta} W_2^2(\mu_t, \nu) + \mathcal{G}(\nu),\ \mathcal{G} : \mathcal{P}_2^r(X) \to \mathbb{R}\,. \tag{3.1.7}$$

Here the functional $\mathcal{G}(\nu) = \mathcal{F}(\nu) + \beta^{-1}\mathcal{E}(\nu)$ consists of two parts, a potential functional $\mathcal{F}$ and an entropy functional $\mathcal{E}$

$$\begin{aligned}
\mathcal{F}(\nu) &:= \int f \,\mathrm{d}\nu = \int f(x) p_\nu(x) \,\mathrm{d}x\ , \\
\mathcal{E}(\nu) &:= \int \log\left(\frac{\mathrm{d}\nu}{\mathrm{d}x}\right) \mathrm{d}\nu = \int p_\nu(x) \log p_\nu(x) \,\mathrm{d}x\ .
\end{aligned}$$

3.2 Denoising and Localization: Score and Curvature

3.2.1 Score Function: Denoising and Optimal Transport

The score function arises naturally in the Fokker-Planck PDE (3.1.6). For a valid probability density ρ, the vector field $\nabla \log \rho : X \to \mathbb{R}^d$ defines the score function.

Definition 3.2.1 (Score Function)**.** For a measure $\nu \in \mathcal{P}_2^r(X)$, denote its density with respect to the Lebesgue measure as p_ν where $\nu = p_\nu \cdot \mathcal{L}^d$. The score function is $x \mapsto \nabla \log p_\nu(x)$.

It turns out, even for non-vanishing η, the score function $\nabla \log \rho_{t+\eta}$ induces an optimal plan to localize and denoise from $\mu_{t+\eta}$, defined in (3.1.7), back to μ_t, in the sense of optimal transport (OT). In the $\eta \to 0$ case, the score induced by the OT map was studied in Chen et al. [2024], Song et al. [2020], Jordan et al. [1998]. We first introduce the OT map.

Definition 3.2.2 (OT Map, [Brenier, 1987])**.** For $\mu, \nu \in \mathcal{P}_2^r(X)$, there exists a unique optimal transport map $\mathbf{t}_\nu^\mu : X \to X$ that solves the Monge problem

$$\mathbf{t}_\nu^\mu = \operatorname*{arg\,min}_{\mathbf{t} : \mathbf{t}_\# \nu = \mu} \left(\int \|y - \mathbf{t}(y)\|^2 \,\mathrm{d}\nu(y) \right)^{1/2}\ .$$

and attains the minimum of the Kantorovich problem

$$W_2(\mu,\nu) = \left(\int \|\mathbf{i} - \mathbf{t}_\nu^\mu\|^2 \mathrm{d}\nu \right)^{1/2} .$$

Proposition 3.2.3 (Score Function and Backward OT Map)**.** *Consider the Wasserstein gradient descent as in* (3.1.7) *with a lower-semicontinuous* $\mathcal{G} = \mathcal{F} + \beta^{-1}\mathcal{E}$*. Assume that there exists* $\eta_\star > 0$*, such that for all* $\eta \in (0, \eta_\star)$*,* $\mu_{t+\eta}$ *in* (3.1.7) *admits a well-defined minimizer. Then for any* $\eta \in (0, \eta_\star)$*, the optimal transport map* $\mathbf{t}_{\mu_{t+\eta}}^{\mu_t}$ *as in Definition 3.2.2 takes the form,*

$$\frac{1}{\eta}(\mathbf{t}_{\mu_{t+\eta}}^{\mu_t} - \mathbf{i})(x) = \nabla f(x) + \beta^{-1}\nabla \log p_{\mu_{t+\eta}}(x), \text{ for } \mu_{t+\eta}\text{-a.e. } x \in \mathbb{R}^d .$$

Recall that discretized Langevin diffusion reads $\mathbf{X}_{t+\eta} = \mathbf{X}_t - \eta\nabla f(\mathbf{X}_t) + \sqrt{2\beta^{-1}\eta}\mathbf{Z}$. Another interpretation of the score function is that it induces a backward denoising step for the diffusion, namely, quantifying the barycenter $\mathbb{E}[\mathbf{X}_t - \eta\nabla f(\mathbf{X}_t) \mid \mathbf{X}_{t+\eta} = y]$. In other words, score estimation can be cast as a prediction problem based on the diffusion process, where one aims to predict $\mathbf{X}_t - \eta\nabla f(\mathbf{X}_t)$ based on $\mathbf{X}_{t+\eta}$ [Saremi and Hyvärinen, 2019].

Proposition 3.2.4 (Score and Backward Denoising)**.** *Consider* $\mathbf{Y} = \mathbf{X} + \sigma\mathbf{Z}$*, where* $\mathbf{X} \sim \mu$ *and* $\mathbf{Z} \sim \mathcal{N}(0, I_d)$ *and* $\mathbf{X}, \mathbf{Z}$ *are independent. Let* $p_{\mathbf{Y}}$ *denote the density function associated with the random variables* $\mathbf{Y}$*. Then*

$$\nabla \log p_{\mathbf{Y}}(y) = -\frac{1}{\sigma^2}\Big\{ y - \mathbb{E}[\mathbf{X}|\mathbf{Y} = y] \Big\} .$$

The score function is the optimal transport map to denoise the diffusion process, but how accurate is the denoising step? Conceptually, the denoising quality depends on the localization $\mathrm{Cov}[\mathbf{X}_t - \eta\nabla f(\mathbf{X}_t) \mid \mathbf{X}_{t+\eta} = y]$. As we shall see next, the localization quality of the score function as a backward denoising step depends on the curvature function, $x \mapsto \nabla^2 \log p_\nu(x)$.

3.2.2 Curvature Function: Backward Localization

The curvature function, namely, the derivative of the score function, governs whether the backward denoising step is localized. The following proposition describes the variability of the backward denoising, $\mathrm{Cov}[\mathbf{X}_t - \eta\nabla f(\mathbf{X}_t) \mid \mathbf{X}_{t+\eta} = y]$. Intuitively, a large positive curvature of the log density function results in a large conditional covariance and in turn, makes the denoising process hard.

Proposition 3.2.5 (Curvature and Localization)**.** *Consider* $\mathbf{Y} = \mathbf{X} + \sigma\mathbf{Z}$*, where* $\mathbf{X} \sim \mu$ *and* $\mathbf{Z} \sim \mathcal{N}(0, I_d)$ *and* $\mathbf{X}, \mathbf{Z}$ *are independent. Let* $p_{\mathbf{Y}}$ *denote the density function associated with the random variables* $\mathbf{Y}$*. Then*

$$\nabla^2 \log p_{\mathbf{Y}}(y) = -\frac{1}{\sigma^2}\Big\{ I_d - \mathrm{Cov}[\tfrac{\mathbf{X}}{\sigma}|\mathbf{Y} = y]\Big\} .$$

The validity of the diffusion-then-denoise process depends on the spectrum of the conditional covariance function. Motivated by this, we shall define a multi-scale complexity measure in Section 3.4. Before concluding this section, we show that the average-case curvature is always negative, although the worst-case curvature is positive for non-log-concave measures. In other words, average curvature $\mathrm{tr}[\nabla^2 \log p_\nu(y)]$ weighed by $p_\nu(y)$, is strictly negative for any ν. This will be useful later.

Proposition 3.2.6 (Curvature and Score)**.**

$$\int \mathrm{tr}[\nabla^2 \log p_\nu(x)]\mathrm{d}\nu(x) = -\int \|\nabla \log p_\nu(x)\|^2 \mathrm{d}\nu(x) .$$

An immediate implication is that the average localization radius is bounded.

$$\mathbb{E}\| \mathrm{Cov}[\tfrac{\mathbf{X}}{\sigma}|\mathbf{Y}]\|_{\mathrm{op}} \le \mathbb{E}\,\mathrm{tr}\left[\mathrm{Cov}[\tfrac{\mathbf{X}}{\sigma}|\mathbf{Y}]\right] \le d .$$

As we shall see in Definition 3.4.2 in Section 3.4.2, the tail behavior of $\| \mathrm{Cov}[\frac{\mathbf{X}}{\sigma}|\mathbf{Y}]\|_{\mathrm{op}}$ governs

the complexity of the diffuse-then-denoise process.

3.3 Forward Contraction and Backward Expansion

In this section, we study how the curvature governs the effectiveness of the diffuse-then-denoise process in diffusion models.

3.3.1 Contraction of Forward Chain

Consider two forward diffusion chains μ_t, ν_t at a given time t, with an infinitesimal time increment η. Recall the Fokker-Planck PDE in (3.1.6); the measures are implemented by

$$\begin{aligned} \mu_{t+\eta} &:= \mathbf{f}^{\mu,\eta}_{\#}\mu_t, \text{ where } \mathbf{f}^{\mu,\eta} = \mathbf{i} - \eta(\nabla f + \beta^{-1}\nabla \log p_{\mu_t}) \,, \\ \nu_{t+\eta} &:= \mathbf{f}^{\nu,\eta}_{\#}\nu_t, \text{ where } \mathbf{f}^{\nu,\eta} = \mathbf{i} - \eta(\nabla f + \beta^{-1}\nabla \log p_{\nu_t}) \,. \end{aligned} \tag{3.3.1}$$

Again, these are the Euler discretizations of the Wasserstein gradient flow as in (3.1.7), with functional $\mathcal{G}(\nu) = \mathcal{F}(\nu) + \beta^{-1}\mathcal{E}(\nu)$.

Theorem 3.3.1 (Forward Contraction)**.** *Assume for some $\lambda \in \mathbb{R}_+$, $\nabla^2 f(x) \succeq \lambda \cdot I_d$, $\forall x \in X$. For any $\mu_t \neq \nu_t \in \mathcal{P}_2^r(X)$ and any $\beta \in \mathbb{R}_+$, the one-step forward process* (3.3.1) *satisfies*

$$\limsup_{\eta \to 0} \frac{1}{\eta} \frac{W_2^2(\mu_{t+\eta}, \nu_{t+\eta}) - W_2^2(\mu_t, \nu_t)}{W_2^2(\mu_t, \nu_t)} \leq -2\lambda \,.$$

Remark 3.3.2. The above Theorem 3.3.1 shows that as $\eta \to 0$, for any μ_t, ν_t

$$\frac{W_2(\mu_{t+\eta}, \nu_{t+\eta})}{W_2(\mu_t, \nu_t)} \leq 1 - \eta\lambda \,.$$

With the choice $\nu_t = \nu_\star$, the invariant measure, we have $\nu_{t+\eta} = \nu_\star$, and thus

$$\frac{W_2(\mu_{t+\eta}, \nu_\star)}{W_2(\mu_t, \nu_\star)} \leq 1 - \eta\lambda \,.$$

Conceptually, the one-step contraction rate for the forward diffusion process is governed by λ. This contraction rate is sharp and holds equality for some μ, ν's, for example, take μ, ν as Dirac measures.

3.3.2 Expansion of Backward Chain

Recall the backward OT map as in Proposition 3.2.3

$$\mathbf{b}^{\mu,\eta} = \mathbf{i} + \eta(\nabla f + \beta^{-1}\nabla \log p_{\mu_t}) \,. \tag{3.3.2}$$

This backward OT map encapsulates the plan to revert the chain $\mathbf{b}^{\mu,\eta}_{\#}\mu_t \to \mu_{t-\eta}$, which we will formally prove in Theorem 3.3.6 in next section. In this section, we conduct a sensitivity analysis on the backward map: consider a measure ν that is a small perturbation to μ_t, will $\mathbf{b}^{\mu,\eta}_{\#}\nu$ stay close to $\mathbf{b}^{\mu,\eta}_{\#}\mu_t$? This question concerns the expansion rate of the backward chain.

We first derive a warm-up result in the simple case, employing a notion of worst-case curvature. Later, we will generalize the result in Section 3.4, elucidating how an average-case curvature gives rise to a multi-resolution complexity measure that governs the effectiveness of the diffuse-then-denoise process.

Theorem 3.3.3 (Backward Expansion)**.** *Consider $\nabla^2 f(x), \nabla^2 \log p_{\mu_t}(x)$ with bounded eigenvalues over $x \in X$. Assume for some $\kappa \in \mathbb{R}_+$, $\nabla^2 f(x) \preceq \kappa \cdot I_d$, $\forall x \in X$. Assume $\nabla \log p_{\mu_t} \in C_c^\infty(X, X)$ and for some $\zeta \in \mathbb{R}$*

$$\nabla^2 \log p_{\mu_t}(x) \succeq -\zeta \cdot I_d, \ \forall x \in X \,.$$

Then for any $\beta \in \mathbb{R}_+$, the one-step backward map (3.3.2) *satisfies*

$$\limsup_{\eta \to 0} \sup_{\nu \in \mathcal{P}_2^r(X)} \frac{1}{\eta} \frac{W_2^2(\mathbf{b}^{\mu,\eta}_{\#}\mu_t, \mathbf{b}^{\mu,\eta}_{\#}\nu) - W_2^2(\mu_t, \nu)}{W_2^2(\mu_t, \nu)} \leq 2(\kappa - \beta^{-1}\zeta) \,.$$

Remark 3.3.4. We emphasize that this theorem operates even when $\zeta < 0$, namely when μ_t

is non-log-concave. The above theorem shows that, for any $\nu \in \mathcal{P}_2^r(X)$, as $\eta \to 0$

$$\frac{W_2(\mathbf{b}_\#^{\mu,\eta}\mu_t, \mathbf{b}_\#^{\mu,\eta}\nu)}{W_2(\mu_t,\nu)} \le 1 + \eta(\kappa - \beta^{-1}\zeta) \; .$$

Conceptually, the one-step expansion rate for the backward OT map is governed by $\kappa - \beta^{-1}\zeta$. We call this an expansion because it is positive when β is large enough, regardless of the sign of ζ.

This backward expansion is inevitable, even for certain log-concave μ_t. As a simple corollary of Theorem 3.3.3, we show the expansion upper bound is tight.

Corollary 3.3.5 (Lower Bound)**.** *Assume for some $\kappa, \zeta \in \mathbb{R}_+$,*

$$\nabla^2 f(x) = \kappa \cdot I_d, \;\; \nabla^2 \log p_{\mu_t}(x) = -\zeta \cdot I_d \; .$$

Then for any $\beta > \zeta/\kappa$, the one-step backward map (3.3.2) *satisfies for all ν*

$$\lim_{\eta \to 0} \frac{1}{\eta} \frac{W_2^2(\mathbf{b}_\#^{\mu,\eta}\mu_t, \mathbf{b}_\#^{\mu,\eta}\nu) - W_2^2(\mu_t,\nu)}{W_2^2(\mu_t,\nu)} = 2(\kappa - \beta^{-1}\zeta) > 0 \; .$$

3.3.3 Diffuse-then-Denoise: One-Step Improvement and Chaining

Consider the special case of an Ornstein–Uhlenbeck process where $f(x) = \|x\|^2/2$. Take any two chains μ_t and ν_t, as $\eta \to 0$, then Theorem 3.3.1 claims the forward diffusion satisfies contraction

$$\frac{W_2(\mathbf{f}_\#^{\mu,\eta}\mu_t, \mathbf{f}_\#^{\nu,\eta}\nu_t)}{W_2(\mu_t,\nu_t)} = \frac{W_2(\mu_{t+\eta},\nu_{t+\eta})}{W_2(\mu_t,\nu_t)} \le 1 - \eta + O(\eta^2) \; .$$

The backward denoising in Theorem 3.3.3 satisfies an expansion at most

$$\frac{W_2(\mathbf{b}_\#^{\mu,\eta}\mathbf{f}_\#^{\mu,\eta}\mu_t, \mathbf{b}_\#^{\mu,\eta}\mathbf{f}_\#^{\nu,\eta}\nu_t)}{W_2(\mathbf{f}_\#^{\mu,\eta}\mu_t, \mathbf{f}_\#^{\nu,\eta}\nu_t)} = \frac{W_2(\mathbf{b}_\#^{\mu,\eta}\mu_{t+\eta}, \mathbf{b}_\#^{\mu,\eta}\nu_{t+\eta})}{W_2(\mu_{t+\eta},\nu_{t+\eta})} \le 1 + \eta(1 - \beta^{-1}\zeta) + O(\eta^2) \; .$$

We shall show next in Theorem 3.3.6 that

$$W_2(\mathbf{b}_{\#}^{\mu,\eta}\mathbf{f}_{\#}^{\mu,\eta}\mu_t, \mu_t) = O(\eta^2) \ .$$

Then compared to ν_t, the diffuse-then-denoise version $\mathbf{b}_{\#}^{\mu,\eta}\mathbf{f}_{\#}^{\nu,\eta}\nu_t$ will get closer to μ_t, in the following sense

$$\frac{W_2(\mu_t, \mathbf{b}_{\#}^{\mu,\eta}\mathbf{f}_{\#}^{\nu,\eta}\nu_t)}{W_2(\mu_t, \nu_t)} \leq 1 - \eta\beta^{-1}\zeta + O(\eta^2) \ .$$

The one-step diffuse-then-denoise process $\mathbf{b}^{\mu,\eta} \circ \mathbf{f}^{\nu,\eta}$ has a net contraction $\beta^{-1}\zeta$ when $\zeta > 0$, namely, the forward contraction, offset by the possible backward expansion, presents a net gain in localization.

Theorem 3.3.6. *Define the one-step diffuse-then-denoise*

$$\begin{aligned}
\mathbf{f}^{\mu,\eta} &:= \mathbf{i} - \eta(\nabla f + \beta^{-1}\nabla \log p_\mu), \ \mu_\eta := \mathbf{f}_{\#}^{\mu,\eta}\mu \ , \\
\mathbf{b}^{\mu,\eta} &:= \mathbf{i} + \eta(\nabla f + \beta^{-1}\nabla \log p_{\mu_\eta}) \ .
\end{aligned}$$

Assume $\nabla f, \nabla \log p_\mu, \nabla \log p_{\mu_\eta} \in C_c^\infty(X, X)$*, then*

$$W_2(\mathbf{b}_{\#}^{\mu,\eta}\mathbf{f}_{\#}^{\mu,\eta}\mu, \mu) = O(\eta^2) \ .$$

A direct corollary concerning the diffuse-then-denoise chain now follows from Theorem 3.3.1, 3.3.3, and 3.3.6. We consider finite K-step diffuse-then-denoise chains:

- Forward diffuse: for the μ chain, set $\mu_0 = \mu$, and define $\mathbf{f}_k^{\mu,\eta} := \mathbf{i} - \eta(\nabla f + \beta^{-1}\nabla \log p_{\mu_{(k-1)\eta}})$ and $\mu_{k\eta} := (\mathbf{f}_k^{\mu,\eta})_{\#}\mu_{(k-1)\eta}$, recursively for $k = 1, 2, \ldots, K$. Same for the ν chain.
- Backward denoise: for the ν chain, recursively apply the backward OT map $\mathbf{b}_k^{\mu,\eta} := \mathbf{i} + \eta(\nabla f + \beta^{-1}\nabla \log p_{\mu_{k\eta}})$ to $\nu_{K\eta}$, for $k = K, K-1, \ldots, 1$.

Corollary 3.3.7 (Diffuse-then-Denoise: Chaining). *Let $f(x) = \|x\|^2/2$. Define the diffuse-then-denoise map*

$$\mathbf{f}^{\nu}_{[K]} := \mathbf{f}^{\nu,\eta}_K \circ \cdots \circ \mathbf{f}^{\nu,\eta}_2 \circ \mathbf{f}^{\nu,\eta}_1 ,$$
$$\mathbf{b}^{\mu}_{[K]} := \mathbf{b}^{\mu,\eta}_1 \circ \mathbf{b}^{\mu,\eta}_2 \circ \cdots \circ \mathbf{b}^{\mu,\eta}_K .$$

Assume there exists a sequence of $\zeta_{k\eta} \in \mathbb{R}$ such that

$$\nabla^2 \log p_{\mu_{k\eta}}(x) \preceq -\zeta_{k\eta} \cdot I_d, \ \forall x \in X, \ \forall k = 1, 2, \ldots, K .$$

Then the diffuse-then-denoise process on ν with fixed K steps, denoted as $\left(\mathbf{b}^{\mu}_{[K]} \circ \mathbf{f}^{\nu}_{[K]}\right)_{\#}\nu$, satisfies

$$\frac{W_2^2\left(\mu, \left(\mathbf{b}^{\mu}_{[K]} \circ \mathbf{f}^{\nu}_{[K]}\right)_{\#}\nu\right)}{W_2^2(\mu,\nu)} \leq \exp\left(-2\eta\beta^{-1}\sum_{k=1}^{K}\zeta_{k\eta} + O(\eta^2)\right) . \tag{3.3.3}$$

The above Corollary states the denoising quality of the diffuse-then-denoise chain is collectively determined by the curvature at each step $\zeta_{k\eta}$. For non-log-concave μ, along the chain, some of the $\zeta_{k\eta}$ will be positive, and some will be negative. Therefore, we need a fine-grained understanding of the curvature at each time scale $t = k\eta$ to understand the whole chain behavior. This calls for a fine-resolution analysis of the curvature of μ_t, going beyond the worst-case curvature, the focus of the next section.

3.4 Beyond Log-Concavity: A Multi-Scale Complexity

In this section, we restrict to the Ornstein-Uhlenbeck process and discover a multi-scale complexity measure that controls the effective contraction/expansion of the diffuse-then-denoise process. The key is that for the Ornstein-Uhlenbeck process, μ_t is a smoothed version of μ at a particular signal-to-noise ratio scale $\mathsf{r} = \mathsf{r}(t)$.

3.4.1 OU Process

Definition 3.4.1 (Ornstein-Uhlenbeck Process)**.** Define the Ornstein-Uhlenbeck Process with initialization $X_0 \sim \mu$, and potential $f(x) = \|x\|^2/2$

$$\mathrm{d}\mathbf{X}_t = -\mathbf{X}_t \mathrm{d}t + \sqrt{2\beta^{-1}}\mathrm{d}\mathbf{B}_t \ .$$

Then the distribution of $\mathbf{X}_t, \forall t \in \mathbb{R}_+$ admits the representation

$$\mathbf{X}_t \overset{\mathcal{L}}{\sim} e^{-t}\mathbf{X}_0 + \sqrt{\beta^{-1}(1-e^{-2t})}\mathbf{Z},\ \mathbf{Z} \sim \mathcal{N}(0, I_d) \ . \tag{3.4.1}$$

For any measure $\mathbf{X} \sim \mu \in \mathcal{P}_2(X)$, the above representation motivates us to consider a sequence of problems indexed by a multi-scale signal-to-noise ratio. Recall Proposition 3.2.5,

$$\begin{aligned} &\mathsf{r}(t) := \frac{e^{-t}}{\sqrt{\beta^{-1}(1-e^{-2t})}},\ \mathsf{s}(t) := \sqrt{\beta^{-1}(1-e^{-2t})} \ , \\ &\nabla^2 \log p_{\mu_t}(x) = -\frac{1}{\mathsf{s}^2(t)} \left\{ I_d - \mathrm{Cov}[\mathsf{r}(t)\mathbf{X}_0 | \mathbf{X}_t = x] \right\} \ . \end{aligned} \tag{3.4.2}$$

The curvature at each scale quantifies and, collectively, determines the measure's complexity.

If μ is a log-concave measure, then by Prékopa-Leindler Inequality, μ_t's are convolutions of a log-concave measure with a Gaussian measure, and therefore, log-concave. Hence, for any t, we know that $\nabla^2 \log p_{\mu_t}(x) \preceq 0$. As a direct consequence of Equation (3.3.3), we know that the diffuse-then-denoise chain is effective.

However, for non-log-concave measure μ, the $\nabla^2 \log p_{\mu_t}(x)$ may have positive eigenvalues. Given Equation (3.4.2), studying the positive eigenvalues of $\nabla^2 \log p_{\mu_t}$ is equivalent to understanding the tail behavior of the localization quantity $\mathrm{Cov}[\mathsf{r}(t)\mathbf{X}_0 | \mathbf{X}_t]$, which motivates the following section.

3.4.2 Multi-Scale Complexity: Non-Log-Concavity and Survival Function

We define a sequence of functions at various signal-to-noise ratio (SNR) scales. In a nutshell, the different scales of SNR probe the tail behavior of the localization quantity, $\mathrm{Cov}[\mathsf{r}(t)\mathbf{X}_0|\mathbf{X}_t]$. In turn, the localization quantity determines the landscape of the curvature of the measure μ_t, a smoothed version of the original measure μ at a given SNR scale $\mathsf{r}(t)$, as shown in Equations (3.4.1)-(3.4.2). This multi-scale corruption also occurred in stochastic localization; see Montanari [2023] for the connection between stochastic localization and diffusions. We give precise complexity measures of the denoising problem, cast in quantities determined by the survival function of the random variable $\mathrm{Cov}[\mathsf{r}(t)\mathbf{X}_0|\mathbf{X}_t]$. As a result, these complexity measures illustrate the effective contraction or expansion rate for the diffuse-then-denoise process, covering the case of non-log-concave μ's.

Two direct consequences of the newly proposed complexity measures follow: (1) It gives rise to a fine-grained analysis for the diffuse-then-denoise process at any SNR scale, for any non-log-concave measures, to be shown in Section 3.4.3; (2) It motivates a simulation-based numerical tool to visualize the bottleneck SNR scale for the denoising problem, as we shall demonstrate in Section 3.5. Several curious phenomena are unveiled and rationalized by the new multi-scale complexity.

Definition 3.4.2 (SNR, Localization and Survival Function)**.** Given an initial measure $\mu \in \mathcal{P}_2(\mathbb{R}^d)$, we define, for any SNR $\mathsf{r} \in \mathbb{R}_{\geq 0}$

$$\mathbf{Y}_{\mathsf{r}} := \mathsf{r}\mathbf{X} + \mathbf{Z},\ (\mathbf{X}, \mathbf{Z}) \sim \mu \otimes \mathcal{N}(0, I_d)$$

where $\mathbf{X} \sim \mu$ and $\mathbf{Z} \sim \mathcal{N}(0, I_d)$ are independent.

Define the localization function and the associated random variable

$$L_{\mathsf{r}}(y) = \|\,\mathrm{Cov}[\mathsf{r}\mathbf{X}|\mathbf{Y}_{\mathsf{r}} = y]\|_{\mathrm{op}}\ ,\ \text{and}\ \mathbf{L}_{\mathsf{r}} = \|\,\mathrm{Cov}[\mathsf{r}\mathbf{X}|\mathbf{Y}_{\mathsf{r}}]\|_{\mathrm{op}}\ ,$$

and denote the survival function of $\mathbf{L}_{\mathsf{r}}$ as $s_{\mathsf{r}}(\cdot) : \mathbb{R}_+ \to [0,1]$

$$s_{\mathsf{r}}(u) := \mathbb{P}(\mathbf{L}_{\mathsf{r}} > u) \ .$$

A few remarks follow from this definition: (i) When the measure $\mu = \mathcal{L}(\mathbf{X})$ is log-concave, then by Prékopa-Leindler Inequality, for all r, the measure $\mathcal{L}(\mathbf{Y}_{\mathsf{r}})$ is log-concave as well. Proposition 3.2.5 implies that the random variable $\mathbf{L}_{\mathsf{r}} \leq 1$. Therefore, the survival function $s_{\mathsf{r}}(1) = 0$. (ii) When the measure $\mu = \mathcal{L}(\mathbf{X})$ is non-log-concave, then the measure $\mathcal{L}(\mathbf{Y}_{\mathsf{r}})$ will be non-log-concave for some r. For these $\mathbf{r}$'s, we know that $\exists y,\ L_{\mathsf{r}}(y) > 1$ and in turn implies $s_{\mathsf{r}}(1) = \mathbb{P}(\mathbf{L}_{\mathsf{r}} > 1) > 0$.

This property is crucial and distinguishes the difficult settings for the backward transport. The validity and effectiveness of reverting the diffusion model depends on the integrated tail of the survival function $s_{\mathsf{r}}(u), u \in [0,1)$, which we define now.

Definition 3.4.3 (Multi-Scale Complexity)**.** For $\delta \in (0,1]$, define

$$h_\mu(\delta, \mathsf{r}) := \int_{1-\delta}^{\infty} s_{\mathsf{r}}(u)\mathrm{d}u \ , \quad m_\mu(\delta, \mathsf{r}) := \frac{h_\mu(\delta, \mathsf{r})}{\delta} \ .$$

Define

$$\delta^*(\mathsf{r}) = \max \left\{ \zeta \in [0,1] : \zeta \in \operatorname*{arg\,min}_{\delta \in [0,1]} \ m_\mu(\delta, \mathsf{r}) \right\} \ ,$$

and the minimal value

$$m^*(\mathsf{r}) := \min_{\delta \in [0,1]} \ m_\mu(\delta, \mathsf{r}) \ .$$

A few observations follow for the $m_\mu(\cdot, \mathsf{r}) : \delta \mapsto h_\mu(\delta, \mathsf{r})/\delta$ function, which ensures $\delta^\star(\mathsf{r})$ and $m^\star(\mathsf{r})$ are well-defined.

Proposition 3.4.4. *The following properties for $m_\mu(\cdot, \mathsf{r})$ hold:*

1. *Log-concave case: If $s_{\mathsf{r}}(1) = 0$, then $m_\mu(\cdot, \mathsf{r})$ is a non-decreasing function in $\delta \in [0, 1]$;*

2. *Non-log-concave case: If $s_{\mathsf{r}}(1) > 0$, then $m_\mu(\cdot, \mathsf{r})$ is either (i) non-increasing in $\delta \in [0, 1]$, or (ii) U-shaped in $\delta \in [0, 1]$, namely first non-increasing then non-decreasing.*

At a high level, if for some $\delta \in [0, 1]$ the integrated tail $\int_{1-\delta}^{\infty} s_{\mathsf{r}}(u)\mathrm{d}u$ is small, then the effective contraction of the diffuse-then-denoise process will be governed by this δ. By Proposition 3.4.4, $m^*(\mathsf{r}) \geq 0$ is well defined and, as we shall see in the next section, this will quantify the type of perturbation that the diffuse-then-denoise step can tolerate and still effectively contract.

3.4.3 Backward Expansion: Refined Analysis

In this section, we give a proof of why the multi-scale survival function at each SNR schedule $\mathsf{r}(t) = \frac{e^{-t}}{\sqrt{\beta^{-1}(1-e^{-2t})}}$ for $t \in [0, \infty]$ controls the backward expansion of the diffusion-then-denoise model. The result generalizes Theorem 3.3.3 to the case of arbitrary non-log-concave measures.

Definition 3.4.5. For a measure $\mu \in \mathcal{P}_2^r(X)$, define a class of measures in reference to μ,

$$\mathcal{M}(\mu, M) := \{\nu \in \mathcal{P}_2^r(X) \ : \ \frac{\sup_{x \in \mathrm{Dom}(\mu)} \|(\mathbf{t}_\mu^\nu - \mathbf{i})(x)\|^2}{\int \|(\mathbf{t}_\mu^\nu - \mathbf{i})(x)\|^2 \mathrm{d}\mu} \leq M\} \ .$$

If $M = 1$, the set $\mathcal{M}(\mu, M)$ consists of simple measures that are a location shift of μ. If $M = \infty$, the set $\mathcal{M}(\mu, M) = \mathcal{P}_2^r(X)$ all Wasserstein space. Here M controls the richness of perturbations around μ.

Theorem 3.4.6 (Backward Expansion: Beyond Log-Concavity)**.** *Recall the $h_\mu(\delta, \mathsf{r})$ function in Defintion 3.4.3. Assume $\nabla^2 \log p_{\mu_t}(x)$ has bounded eigenvalues over $x \in X$. Consider the*

backward OT map for the OU process as in Definition 3.4.1, then for any $\delta \in [0,1]$

$$\limsup_{\eta \to 0} \limsup_{\substack{\nu \xrightarrow{W_2} \mu_t \\ \nu \in \mathcal{M}(\mu_t, M)}} \frac{1}{\eta} \frac{W_2^2(\mathbf{b}_\#^{\mu,\eta}\mu_t, \mathbf{b}_\#^{\mu,\eta}\nu) - W_2^2(\mu_t, \nu)}{W_2^2(\mu_t, \nu)} \leq 2 - \tfrac{2}{1-e^{-2t}}\left[\delta - M \cdot h_\mu(\delta, \mathsf{r}(t))\right] .$$

Here the SNR schedule $\mathsf{r}(t)$ *is defined in* (3.4.2).

To build intuition for this result, we will isolate the following term,

$$\zeta_M^*(t) = \sup_{\delta \in [0,1]} \tfrac{1}{1-e^{-2t}}\left[\delta - M \cdot h_\mu(\delta, \mathsf{r}(t))\right] .$$

We shall extensively explore how the multi-scale complexity affects this effective curvature complexity $\zeta_M^*(t)$ with several non-log-concave examples in the next section. The curious reader may wonder whether this multi-scale complexity recovers the result obtained in the previous section for the log-concave case. The answer is yes.

In the log-concave case, namely, there exists $\zeta \geq 0$, $\nabla^2 \log p_{\mu_t}(x) \preceq -\zeta \cdot I_d$. For $\tilde{\zeta}(t) = (1-e^{-2t})\zeta$ and $\mathsf{r} = \mathsf{r}(t)$, we have $\nabla^2 \log p_{\mathbf{Y}_\mathsf{r}}(y) \preceq -\tilde{\zeta} \cdot I_d$. In this case, $L_\mathsf{r}(y) \leq 1 - \tilde{\zeta}$ and $s_\mathsf{r}(1-\tilde{\zeta}) = 0$. Then, $m_\mu(\delta, \mathsf{r}) = 0$ for all $\delta \in [0, \tilde{\zeta}]$, and

$$\begin{aligned} &m^*(\mathsf{r}) = 0, \quad \delta^*(\mathsf{r}) = \tilde{\zeta}, \\ &\zeta_\infty^*(t) = \lim_{M \to \infty} \frac{1}{1-e^{-2t}}\left[\delta^*(\mathsf{r}) - M \cdot h_\mu(\delta^*(\mathsf{r}), \mathsf{r})\right] = \frac{\tilde{\zeta}}{1-e^{-2t}} = \zeta. \end{aligned}$$

In the non-log-concave case, for any $M < \frac{1}{m^*(\mathsf{r})}$, it is possible to obtain $\zeta_M^*(t) > 0$, and thus, a net contraction for the diffuse-then-denoise process. We will explore this extensively in the next section and delineate the behavior of $\zeta_M^*(t)$ for a host of fundamental non-log-concave distributions.

3.5 Examples

Theorem 3.4.6 depends on an integrated survival function $h_\mu(\delta, \mathsf{r})$ whose shape is difficult to guess. Luckily, its complete behavior can be captured by understanding the following quantities.

$$s_{\mathsf{r}}(u) = \mathbb{P}(\mathbf{L}_{\mathsf{r}} > u),\ m^*(\mathsf{r}) = \min_{\delta\in[0,1]} m_\mu(\delta, \mathsf{r}),\ \delta^*(\mathsf{r}) = \max\Big\{\zeta \in [0,1] : \zeta \in \arg\min_{\delta\in[0,1]}\ m_\mu(\delta, \mathsf{r})\Big\}.$$

We will methodically calculate and plot these objects for a range of fundamental distributions. Given a target distribution, empirical or with explicit form, we can use a Monte Carlo simulation method to visualize the functions above.

- For any empirical measure μ_0, (3.5.3) defines an expression for the empirical version of $\mathbf{L}_r(y)$.
- One can simulate $\mathbf{Y}_{\mathsf{r}}$ by sampling $\mathsf{r}X + \mathcal{N}(0,1),\ \text{for } X \sim \mu_0$.
- Then one can obtain the empirical survival function, $s_{\mathsf{r}}(u)$. $m^*(\mathsf{r})$, $\delta^*(\mathsf{r})$ and $\zeta^*_M(t)$ follow.

Capturing this behavior at precise time scales will allow us to ascertain (1) when contraction is easy, and if not, (2) at what SNR, r, the process transitions to a non-log-concave setting. In the following, we will restrict ourselves to the OU process, with $\beta = 1$, and in one dimension.

3.5.1 Warm-Up: Log-Concave

Suppose the target measure, μ_0 is log-concave. Via Prékopa-Leindler Inequality, the full chain of measures generated by the OU process admits log-concavity. That is, there exists a sequence of non-negative constants $\zeta(\mathsf{r}) > 0$ such that

$$\nabla^2 \log p_{\mathbf{Y}_{\mathsf{r}}}(y) \preceq -\zeta(\mathsf{r}), \quad \forall \mathsf{r} \geq 0.$$

We know from the previous section that this implies,

$$m^*(\mathsf{r}) = 0, \quad \delta^*(\mathsf{r}) = \zeta(\mathsf{r}), \quad \zeta_\infty^*(t) = \frac{\delta^*(r(t))}{1 - e^{-2t}}.$$

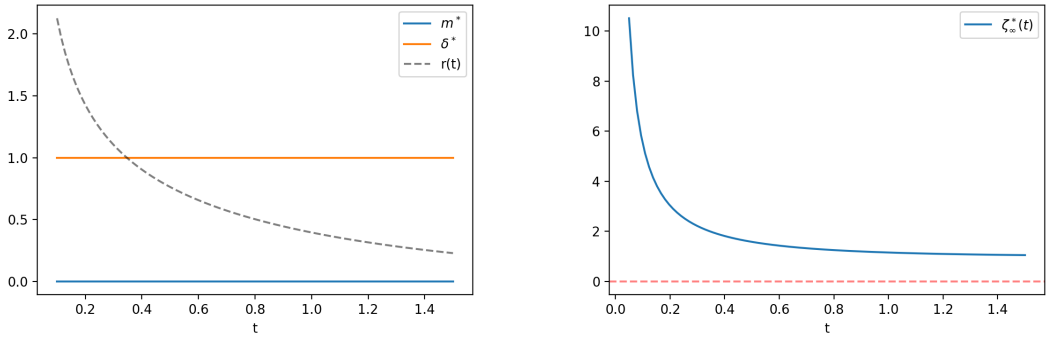
We visualize and validate this result for three log-concave distributions.

Example 3.5.1 (Point Mass). Consider the simplest case where the target measure μ_0 is a Dirac measure, δ_0. We assess the localization of the backward transport by investigating the random covariance, $\mathbf{L}_\mathsf{r}$.

$$-\nabla^2 \log p_{\mathbf{Y}_\mathsf{r}}(y) = 1, \quad \implies \mathbf{L}_\mathsf{r} \equiv 0.$$

In this simple case, $\mathbf{L}_\mathsf{r}$ is a point mass at 0. The backward transport localizes completely. It follows that $s_\mathsf{r}(u) = 0$. Therefore,

$$m^*(\mathsf{r}) = 0, \quad \delta^*(\mathsf{r}) = 1, \quad \zeta_\infty^*(t) = \frac{1}{1 - e^{-2t}}.$$



(a) *(b)*

Figure 3.3: (a) $m^*(\mathsf{r}), \delta^*(\mathsf{r})$. (b) $\zeta^*(t)$.

Example 3.5.2 (Normal). Consider now that μ_0 is a normal distribution $\mathcal{N}(m, \sigma^2)$. Again, we calculate the localization quantity,

$$-\nabla^2 \log p_{\mathbf{Y}_\mathsf{r}}(y) = \frac{1}{\sigma^2\mathsf{r}^2 + 1}, \quad \implies \mathbf{L}_\mathsf{r} = \frac{\sigma^2\mathsf{r}^2}{\sigma^2\mathsf{r}^2 + 1}.$$

$\mathbf{L}_{\mathsf{r}}$ is a point mass with location depending on the SNR, r. In turn, the survival function is not 0, but a step function with an explicit threshold,

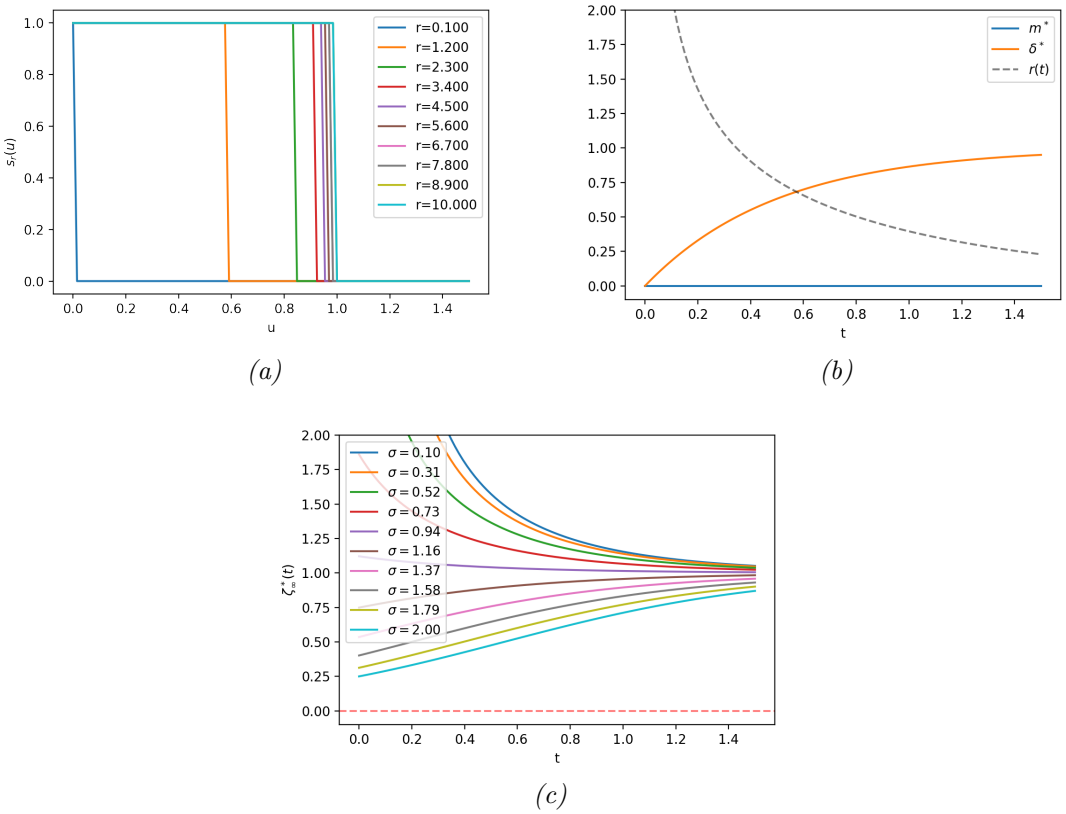
$$s_{\mathsf{r}}(u) = \begin{cases} 1, & \forall u \in [0, \frac{\sigma^2\mathsf{r}^2}{\sigma^2\mathsf{r}^2+1}) \\ 0, & \forall u \in [\frac{\sigma^2\mathsf{r}^2}{\sigma^2\mathsf{r}^2+1}, \infty) \end{cases}$$

We plot this in Figure 3.4 (a). The backward transport localizes completely as $\mathsf{r} \to 0$ (or equivalently $t \to \infty$).

For any r, we can find $\delta \in [0, 1]$ such that $m_\mu(\delta, \mathsf{r}) = 0$. It follows,

$$m^*(\mathsf{r}) = 0, \quad \delta^*(\mathsf{r}) = \frac{1}{\sigma^2\mathsf{r}^2 + 1}, \quad \zeta^*_\infty(t) = \frac{1}{1 + e^{-2t}(\sigma^2 - 1)}.$$

We confirm a variety of σ^2 choices in Figure 3.4 (b) and (c).



(a) *(b)*

(c)

Figure 3.4: (a) $s_{\mathsf{r}}(u)$. (b) $m^*(\mathsf{r}), \delta^*(\mathsf{r})$. (c) $\zeta^*(t)$.

Example 3.5.3 (Uniform). Consider the case when the target distribution is uniform $\mu_0 = \text{Unif}(-1, 1)$. As before, we gauge the localization via the random covariance of the backward transport, $\mathbf{L}_{\mathsf{r}}$.

$$\nabla^2 \log p_{\mathbf{Y}_{\mathsf{r}}}(y) = \Big[\frac{-(y+\mathsf{r})\phi(y+\mathsf{r}) + (y-\mathsf{r})\phi(y-\mathsf{r})}{\Phi(y+\mathsf{r}) - \Phi(y-\mathsf{r})} - \left(\frac{\phi(y+\mathsf{r}) - \phi(y-\mathsf{r})}{\Phi(y+\mathsf{r}) - \Phi(y-\mathsf{r})}\right)^2\Big] ,$$
$$L_{\mathsf{r}}(Y_{\mathsf{r}}) = 1 + \nabla^2 \log p_{Y_{\mathsf{r}}}(Y_{\mathsf{r}}) .$$

Based on this characterization we simulate $s_{\mathsf{r}}(u),\ m^*(\mathsf{r}),\ \delta^*(\mathsf{r}),\ \zeta^*_\infty(t)$. As expected we see,

$$s_{\mathsf{r}}(1) = 0, \quad m^*(\mathsf{r}) = 0, \quad \zeta^*_\infty(t) > 0.$$

We recover the same qualitative behavior as in the preceding examples (point mass, gaussian). However, the fine-grained behavior is remarkably different; see Figure 3.5 (c). Note that though $\zeta^*_\infty(t) > 0$, we have the non-monotonic behavior of the effective contract at different time scales. The slowest effective contraction is happening in some small time scale.

The uniting technicality that allows this is:

$$\mathbf{L}_{\mathsf{r}} \leq 1, \iff s_{\mathsf{r}}(1) = 0, \quad \forall \mathsf{r} \geq 0.$$

By Proposition 3.2.5, this is exactly saying that the curvature, $\nabla^2 \log p_{\mathbf{Y}_{\mathsf{r}}}(y) \leq 0$.

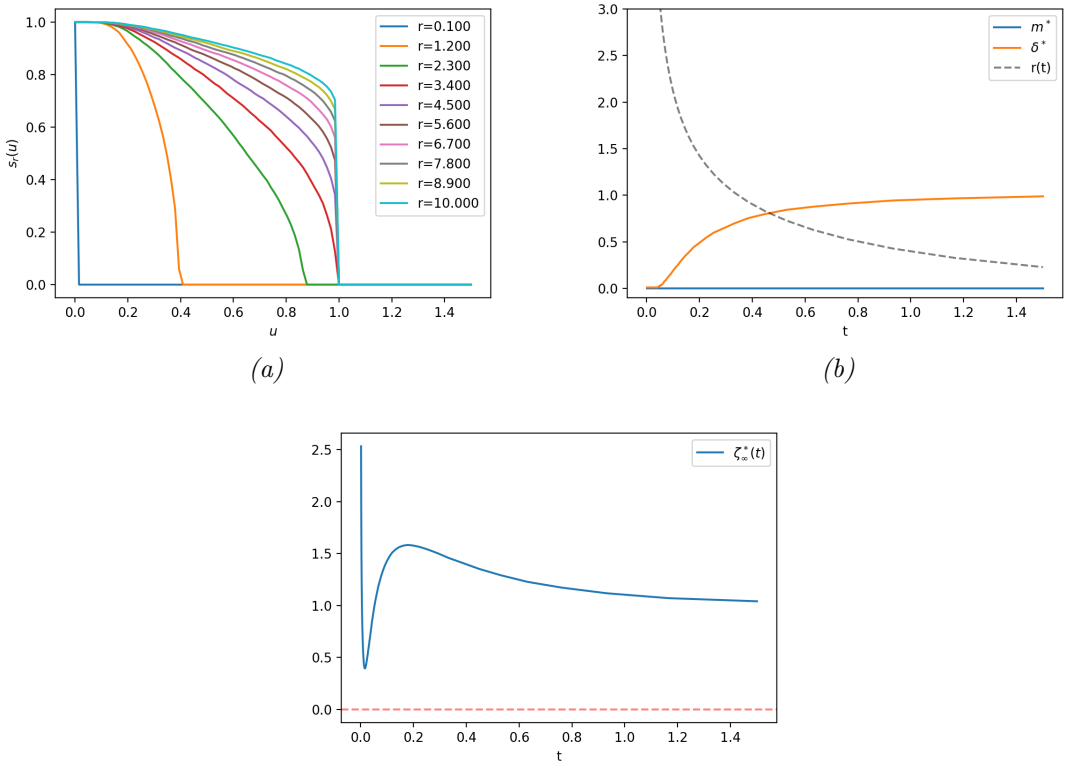


Figure 3.5: (a) $s_{\mathsf{r}}(u)$. (b) $m^*(\mathsf{r}), \delta^*(\mathsf{r})$. (c) $\zeta^*(t)$.

3.5.2 Beyond Log-Concavity

With general non-log-concave μ as initialization, the Ornstein-Uhlenbeck process at time t may not have log-concavity. Indeed, $\nabla^2 \log p_{\mathbf{Y}_{\mathsf{r}}}(y)$ is not uniformly upper bounded by 0, or equivalently, the random covariance $\mathbf{L}_{\mathsf{r}}$ cannot be uniformly bounded from above by 1. We separate from the regime of the previous three examples, all of which exploited this property to show contraction at all scales r.

It is important to understand (i) when the backward chain enters this non-log-concave regime, and (ii) the expansion behavior in this regime. To that end, we make use of the following notation:

$$\mathsf{r}^* = \max\{\mathsf{r} : s_{\mathsf{r}}(1) = 0\}.$$

Example 3.5.4 (Two Point Mass). We consider $X_0 \sim \frac{1}{2}\delta_{-\mu} + \frac{1}{2}\delta_{\mu}$ for some $\mu > 0$. We assess the localization via the random variance, $\mathbf{L}_{\mathsf{r}}$.

$$L_{\mathsf{r}}(y) = (\mathsf{r}\mu)^2\Big[1 - \Big(\frac{\phi(y-\mathsf{r}\mu) - \phi(y+\mathsf{r}\mu)}{\phi(y-\mathsf{r}\mu) + \phi(y+\mathsf{r}\mu)}\Big)^2\Big], \quad \sup_y \; L_{\mathsf{r}}(y) = L_{\mathsf{r}}(0) = (\mathsf{r}\mu)^2.$$

We see immediately, $\mathsf{r}^* = \mu^{-1}$. We can explicitly calculate $s_{\mathsf{r}}(u)$.

$$s_{\mathsf{r}}(u) = \Phi\left(\frac{1}{2(\mathsf{r}\mu)}\log\frac{\mathsf{r}\mu + \sqrt{(\mathsf{r}\mu)^2 - u}}{\mathsf{r}\mu - \sqrt{(\mathsf{r}\mu)^2 - u}} + \mathsf{r}\mu\right) + \Phi\left(\frac{1}{2(\mathsf{r}\mu)}\log\frac{\mathsf{r}\mu + \sqrt{(\mathsf{r}\mu)^2 - u}}{\mathsf{r}\mu - \sqrt{(\mathsf{r}\mu)^2 - u}} - \mathsf{r}\mu\right) - 1 \tag{3.5.1}$$

$$s_{\mathsf{r}}(u) = 0, \quad u \in [(\mathsf{r}\mu)^2, \infty). \tag{3.5.2}$$

(i) ($\mathsf{r} \le \mu^{-1}$): In this setting, we expect log-concave behavior. By (3.5.2),

$$m^*(\mathsf{r}) = 0, \quad \delta^*(\mathsf{r}) = 1 - (\mathsf{r}\mu)^2, \quad \zeta^*_\infty(t) = \frac{1 - (\mathsf{r}(t)\mu)^2}{1 - e^{-2t}} \overset{t\to\infty}{\to} 1.$$

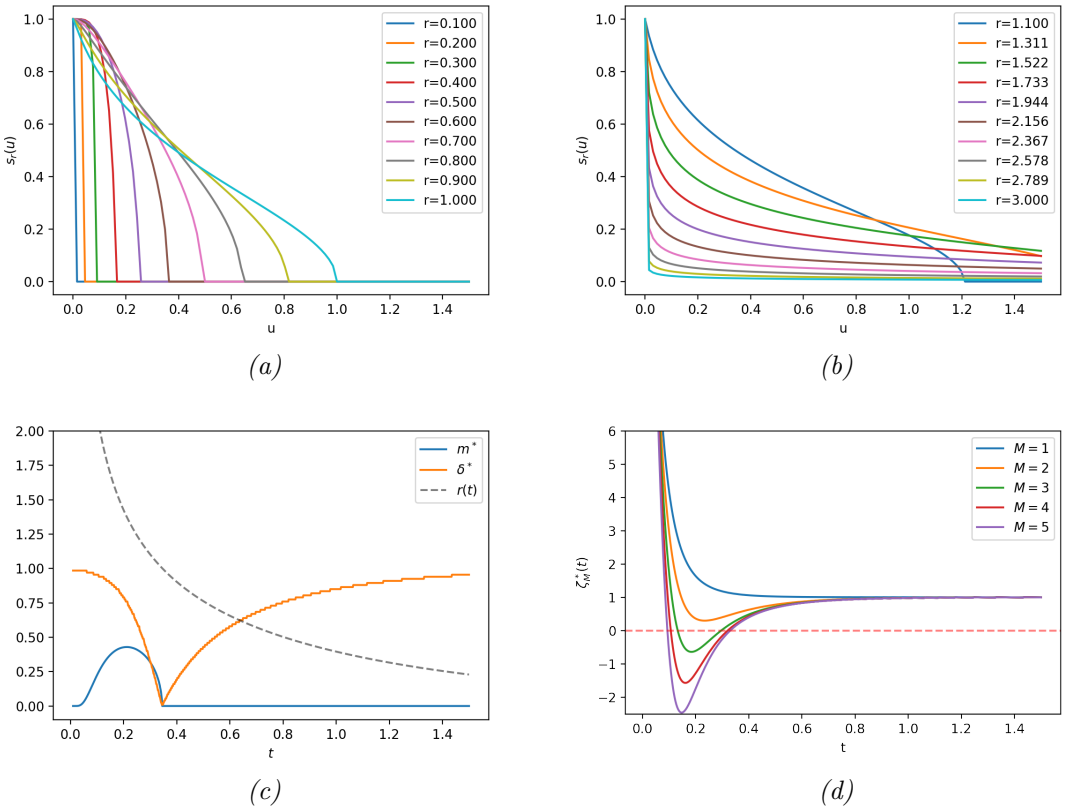
(ii) ($\mathsf{r} \to \infty$): For extremely high SNR, we can see $s_{\mathsf{r}}(u) \to 0$. This phenomenon is qualitatively the same as log-concavity and can be rationalized as essentially recovering the single-point mass behavior in the limit. In particular,

$$m^*(\mathsf{r}) \to 0, \quad \delta^*(\mathsf{r}) \to 1, \quad \zeta^*_\infty(t) \to \infty.$$

(iii) Contraction should be hardest at mid-range SNR, where we no longer have log-concavity. The behavior is simulated and displayed in Figure 3.6 (c). In this setting, we can tolerate a complexity $M \le \frac{1}{m^*(\mathsf{r})} < \infty$ in order for non-expansion $\zeta^*_M(t) \ge 0$. This is visualized in Figure 3.6 (d). Note the non-monotonic behavior of the curvature complexity $\zeta^*_M(t)$: there seems to be a mid-range of time where the diffuse-then-denoise chain could expand, when the type of perturbation is complex with large M.

The conclusion is not pessimistic. For most of the process, regardless of M, $\zeta^*_M(t)$ is

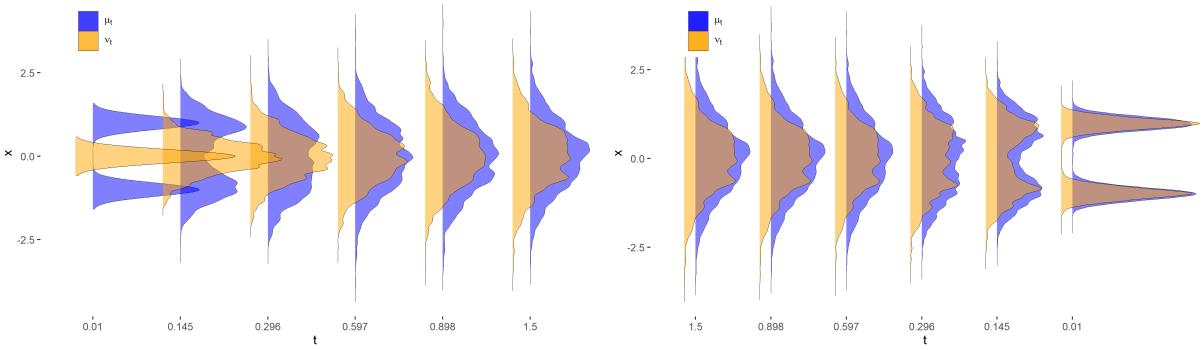
positive and contraction occurs. We remind the reader that M is a proposed upper bound in Definition 3.4.5. In our numerical examples, this is typically small. Refer to Figure 3.7 to see the success of diffuse-then-denoise in this case.



(a) *(b)*

(c) *(d)*

Figure 3.6: (a) $s_{\mathsf{r}}(u)$ Low SNR. (b) $s_{\mathsf{r}}(u)$ High SNR. (c) $m^*(\mathsf{r}), \delta^*(\mathsf{r})$. (d) $\zeta_M^*(t)$.



(a) Forward Diffusion via OU *(b) Backward Transport via OT*

Figure 3.7: (a) Forward diffusion initialized at $\mu_0 = 0.5\delta_1 + 0.5\delta_{-1}$, $\nu_0 = \delta_0$. $T = 100$, $\eta = 0.01$. (b) Backward transport initialized at μ_T, ν_T, and applying the backward OT map $\mathbf{b}^{\mu,\eta}$.

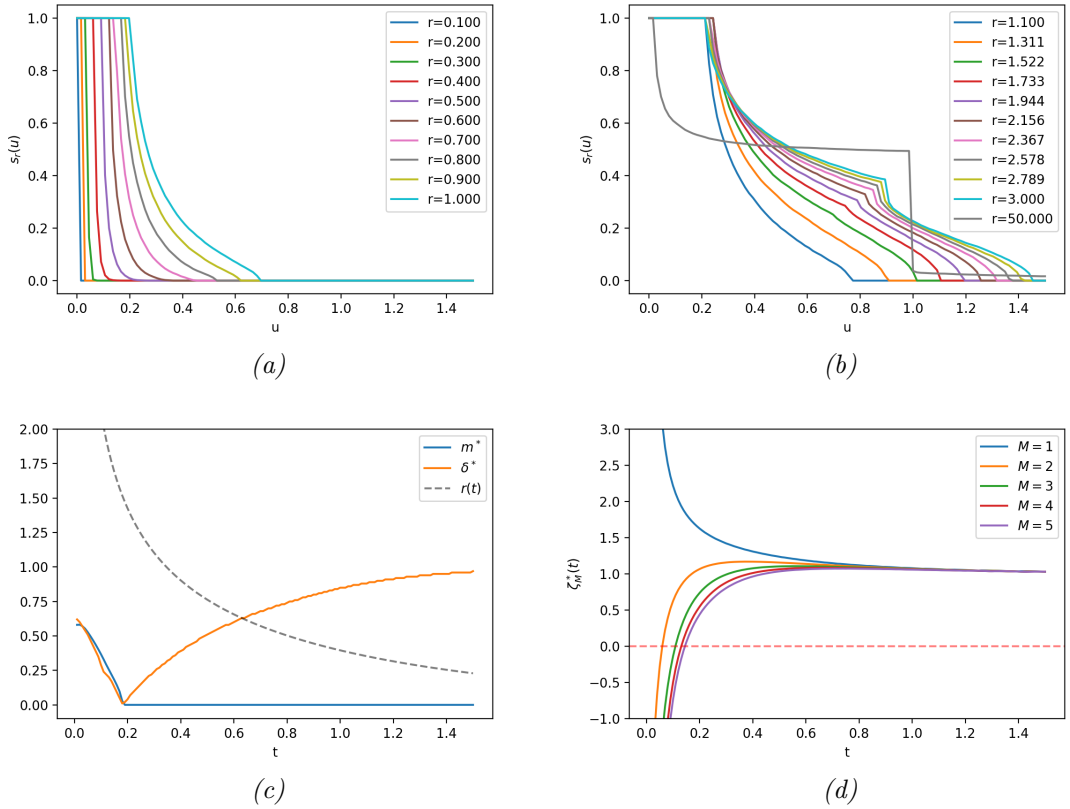


Figure 3.8: (a) $s_{\mathsf{r}}(u)$ Low SNR. (b) $s_{\mathsf{r}}(u)$ High SNR. (c) $m^*(\mathsf{r}), \delta^*(\mathsf{r})$. (d) $\zeta^*_M(t)$.

Example 3.5.5 (Mixture of Point Mass and Gaussian). We now consider diffuse-then-denoise for the target distribution; $\frac{1}{2}\delta_0 + \frac{1}{2}\mathcal{N}(0,1)$. Let $v = \sqrt{\mathsf{r}^2+1}$. We first calculate the covariance,

$$L_{\mathsf{r}}(y) = 1 + \frac{v^{-3}\left(y^2/v^2 - 1\right)\phi\left(y/v\right) + (y^2-1)\phi(y)}{v\phi\left(y/v\right) + \phi(y)} - \left(\frac{v^{-3}y\phi\left(y/v\right) + y\phi(y)}{v\phi\left(y/v\right) + \phi(y)}\right)^2.$$

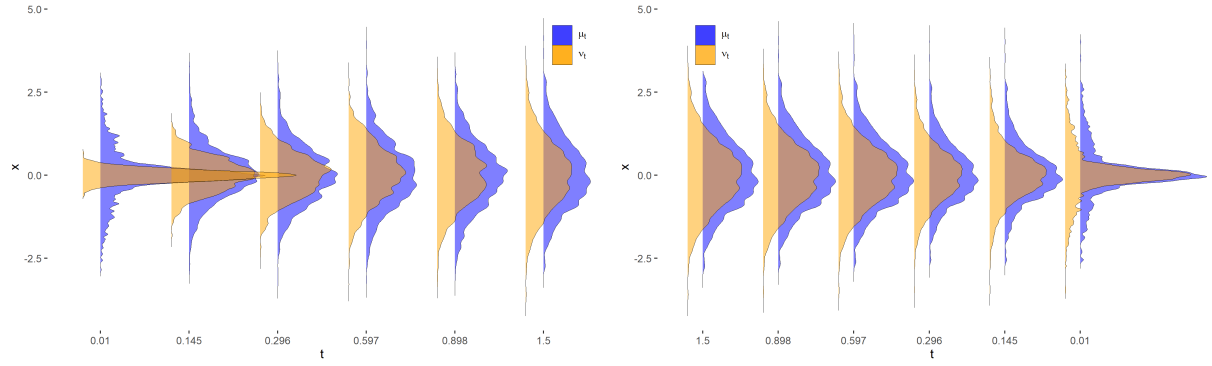
We simulate the survival function $s_{\mathsf{r}}(u) = \mathbb{P}(\mathbf{L}_{\mathsf{r}} > u)$ via Monte Carlo, see Figures 3.8 (a) and (b). A lack of log-concavity is clear for high SNR.

(i) ($\mathsf{r} \le \mathsf{r}^*$): We expect $m^*(\mathsf{r}) = 0$. While we don't have an explicit form for r^*, simulations in Figure 3.8 validate this. See also Monte Carlo-based simulations for δ^*, ζ^*_∞.

(ii) ($\mathsf{r} > \mathsf{r}^*$): We move away from the log-concave regime for large SNR (small t). We conjecture that as $r \to \infty$, $m^*(\mathsf{r}) \to \frac{1}{2}$. To see this, note that $s_{\mathsf{r}}(t)$ approaches a step function

form with a step at 0.5. In this setting, the behavior of ζ_M^* is captured in Figure 3.8 (d). We note that regardless of M, the range of time when the process effectively expands is minimal compared to the full process.

Refer to Figure 3.9 to see the success of diffuse-then-denoise in this case.



(a) Forward Diffusion via OU *(b) Backward Transport via OT*

Figure 3.9: (a) Forward diffusion initialized at $\mu_0 = 0.5\delta_0 + 0.5\mathcal{N}(0,1)$, $\nu_0 = \delta_0$. $T = 100$, $\eta = 0.01$. *(b) Backward transport initialized at* μ_T, ν_T, *and applying the backward OT map* $\mathbf{b}^{\mu,\eta}$.

Example 3.5.6 (Mixture of Gaussian, Heterogeneous Variance). Consider the general Gaussian Mixture given by,

$$X_0 \sim \sum_{i=1}^{m} p_i \cdot \mathcal{N}(\mu_i, \sigma_i^2)$$

In Figure 3.10, we consider a specific formulation:

$$0.1\mathcal{N}(-4,1) + 0.2\mathcal{N}(-2,0.5) + 0.4\mathcal{N}(2,0.5) + 0.3\mathcal{N}(4,1).$$

Defining $v_i = \sqrt{\mathsf{r}^2\sigma_i^2 + 1}$, $t_i = {v_i}^{-1}(y - \mathsf{r}\mu_i)$, for $i \in [m]$, we derive the conditional covariance:

$$L_{\mathsf{r}}(y) = 1 + \frac{\sum_i p_i v_i^{-3}(t_i^2 - 1)\phi(t_i)}{\sum_i p_i v_i \phi(t_i)} - \left(\frac{-\sum_i p_i v_i^{-2} t_i \phi(t_i)}{\sum_i p_i v_i \phi(t_i)}\right)^2 \tag{3.5.3}$$

The survival function $s_{\mathsf{r}}(u) = \mathbb{P}(\mathbf{L}_{\mathsf{r}} > u)$ is simulated via Monte Carlo and is displayed in

Figures 3.10 (a) and (b). As soon as the SNR is not small, we are no longer in the log-concave setting.

(i) ($\mathsf{r} \leq \mathsf{r}^*$): We expect $m^*(\mathsf{r}) = 0$. Simulations in Figure 3.10 show a flat m^* for sufficiently large r. Even for mid-range SNR, we diverge from this log-concave setting.

(ii) ($\mathsf{r} > \mathsf{r}^*$): The shape of ζ_M^* is captured in Figure 3.10 (d). In contrast to the previous two cases, this region is not so narrow. For a chosen noise schedule, diffuse-then-denoise still seems successful for sampling in this setting, see Figure 3.11.

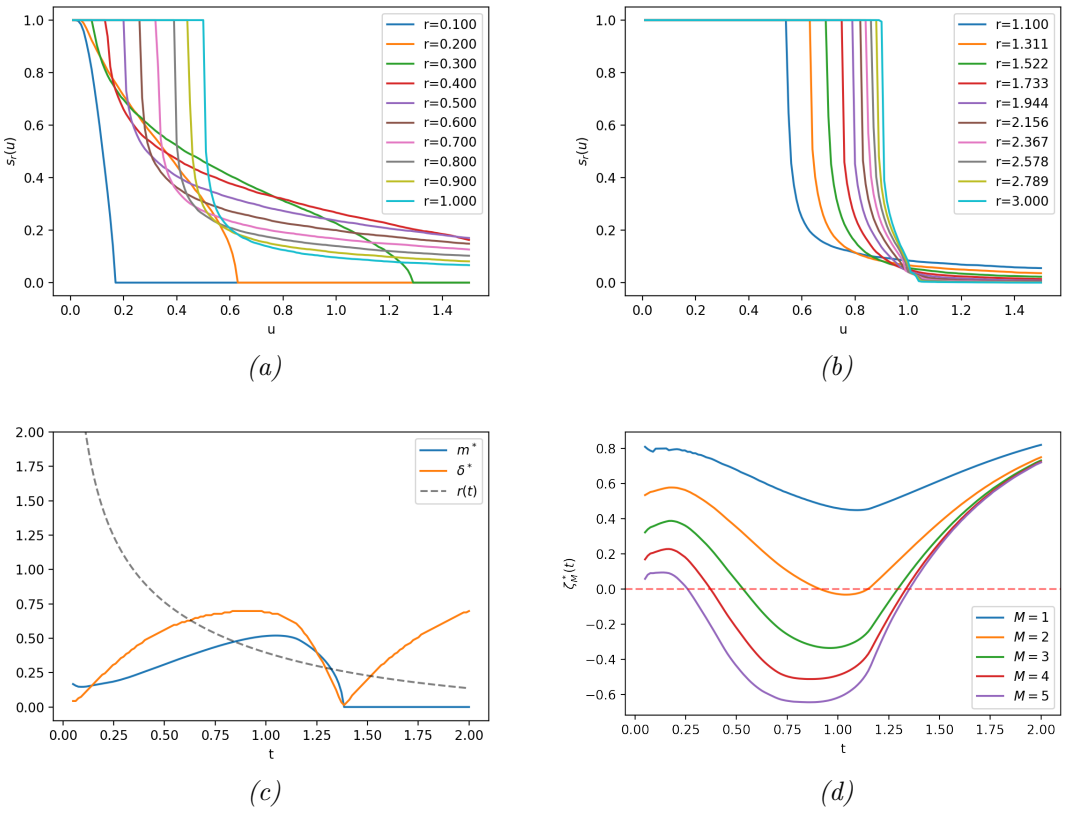
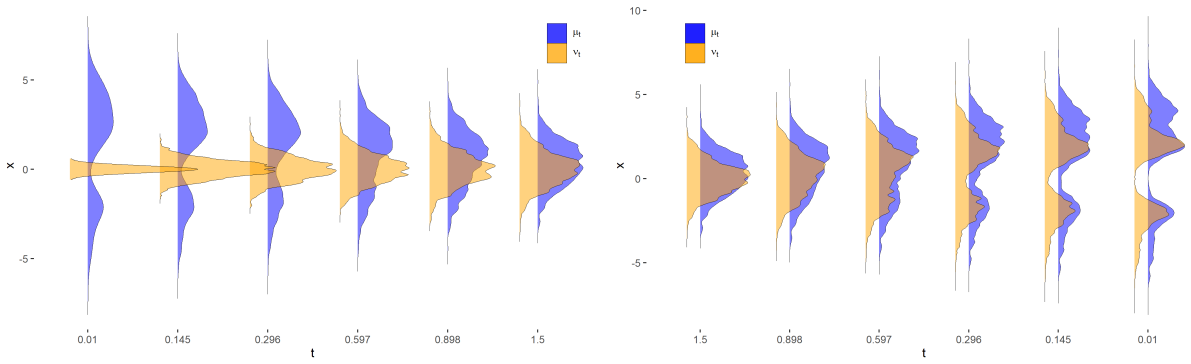


Figure 3.10: (a) $s_{\mathsf{r}}(u)$ Low SNR. (b) $s_{\mathsf{r}}(u)$ High SNR. (c) $m^*(\mathsf{r}), \delta^*(\mathsf{r})$. (d) $\zeta_M^*(t)$.



(a) Forward Diffusion via OU *(b) Backward Transport via OT*

Figure 3.11: (a) Forward diffusion initialized at $\mu_0 = 0.1\mathcal{N}(-4,1) + 0.2\mathcal{N}(-2,0.5) + 0.4\mathcal{N}(2,0.5) + 0.3\mathcal{N}(4,1)$, $\nu_0 = \delta_0$. $T = 100$, $\eta = 0.01$. *(b) Backward transport initialized at* μ_T, ν_T, *and applying the backward OT map* $\mathbf{b}^{\mu,\eta}$.

CHAPTER 4

DISTRIBUTIONAL DENOISING AS LOCAL TRANSPORT

The previous chapter suggests a simple organizing principle: diffusion models make denoising local. The forward process constructs a path of neighboring distributions by adding small amounts of noise over many steps. The reverse problem, transporting from reference to target distribution, is decomposed from a single large recovery problem to a sequence of small ones.

This chapter uses that observation to revisit the classical additive-noise model in the small-noise regime.

$$Y = X + \sigma Z.$$

When σ is small, this model describes the local problem that appears inside a finely discretized diffusion chain. The corrupted distribution P_Y is then only a small deformation of the signal distribution P_X. The reverse operation should therefore preserve most of the observed noisy sample, while adjusting for the local corruption introduced by the noise. It is natural, in this regime, to look for a denoising map close to the identity. In mathematical terms, this means considering maps of the form

$$T(y) = y + \eta v(y), \qquad \eta \text{ small.}$$

This structure is already familiar from score-based denoising. In the notation of the previous chapter, the local denoising map in score-based diffusion models has the approximate form

$$T(y) = y + \eta\{y + \nabla \log q(y)\},$$

where q denotes the density of the noisy observation. Although this map was motivated there through the language of Wasserstein gradient flows and JKO optimality, its form is simple and familiar: it is a small perturbation of the identity, informed by local structure in

the noisy distribution.

In this chapter, we reduce this question to its essential form. Given a noisy distribution P_Y, what structure should the local map T have in order to recover the signal distribution P_X? How should one construct a map whose push-forward of the noisy distribution is close to the signal distribution,

$$T\sharp P_Y \approx P_X?$$

These questions isolate the distributional recovery problem created by diffusion. They lead from score-based denoising to a local transport perspective, in which the reverse step is understood as a map between neighboring probability distributions.

4.1 The Small-Noise Denoising Problem

Consider a single step of a forward noising procedure. In the additive Gaussian model, this step is written as

$$Y = X + \sigma Z, \qquad Z \sim N(0, I_d),$$

where σ is small. Let P denote the distribution of X, let Q denote the distribution of Y, and let q be the density of Y. As $\sigma \to 0$, the corrupted distribution Q approaches the signal distribution P. The reverse problem is therefore local: it asks how to move samples from a small Gaussian perturbation of P back toward P itself.

This is exactly the structure that appears inside a discretized diffusion chain. For example, a local Ornstein–Uhlenbeck step may be written

$$X_{t+\eta} = (1-\eta)X_t + \sqrt{2\eta}\, Z.$$

If $\widetilde{X}_t = (1-\eta)X_t$ and $Y = X_{t+\eta}$, then

$$Y = \widetilde{X}_t + \sqrt{2\eta}\, Z.$$

Thus each fine reverse step is an additive Gaussian denoising problem, with noise level determined by the discretization scale η. The small-noise regime is built into the way the forward path is discretized in a diffusion model.

This local perspective suggests a perturbative form for the reverse map. Since the noisy and less noisy distributions are close, the map should begin at the identity and then add corrections determined by the noisy distribution. It turns out, this intuition is correct. In this chapter we will show how transport maps of the form,

$$\mathrm{T}(y) = y + \eta_1 v_1(y) + \eta_2 v_2(y) + \cdots .$$

are provable improvements over classical denoisers, closely following the work of Liang and co-authors [Liang, 2025a,b].

A change in criterion. Before formalizing this intuition, it is worthwhile to discuss a shift in how denoising maps will be evaluated in this new framework. Traditionally, denoising maps are evaluated through a point-wise criterion. In the empirical-Bayes setting discussed in Chapter 2, the denoiser is chosen to minimize

$$\mathbb{E}\|X - \mathrm{T}(Y)\|^2,$$

When Z is Gaussian, Tweedie's formula provides an explicit form for the Bayes optimal action,

$$\mathrm{T}^\star(y) = y + \sigma^2 \nabla \log q(y).$$

However, now the object of interest is the distribution of $\mathrm{T}\sharp P_Y$. We seek a notion of distributional discrepancy or distance – can we say how closely the signal distribution P_X and the denoised distribution $\mathrm{T}\sharp P_Y$ match? It is known that the optimal transport map T^{ot} transports P_Y to P_X in minimum Wasserstein distance and satisfies the Monge-Ampere

equation

$$p\Big(\ \mathrm{T}^{\mathrm{ot}}(y)\Big)\det\Big(\nabla \mathrm{T}^{\mathrm{ot}}(y)\Big) - q(y) = 0. \tag{4.1.1}$$

Inspired by this, Liang [2025a] motivates an affinity that measures how well $\mathrm{T}\sharp P_Y$ matches P_X, restricted to a compact domain $\Omega \subset \mathbb{R}^d$.

$$d_{\mathrm{MA}}\left(\ \mathrm{T}\sharp P_Y, P_X\right) := \sup_{y\in\Omega} \left| p_Y(y) - p_X(\ \mathrm{T}(y))\det(\nabla\ \mathrm{T}(y))\right|. \tag{4.1.2}$$

In the sequel we will demonstrate how this shift in criterion yields new optimal denoisers– ones we can use in diffusion based sampling schemes. The exposition is intended to develop intuition and understanding, rather than dwell on technical rigor. An interested reader is encouraged to read the literature (Liang [2025a,b]).

4.2 Local Transport Expansions

We now return to the perturbative form suggested above. In the small-noise regime, it is natural to search for a denoising map close to the identity,

$$T(y) = y + \eta_1 v_1(y) + \eta_2 v_2(y) + \cdots .$$

The question is how the vector fields $v_1, v_2, \ldots$ should be chosen. The Monge–Ampère equation (4.1.1) gives a natural target: the optimal transport map from the noisy distribution to the signal distribution. The perturbative question is then whether one can approximate this map order by order, using successive corrections to the identity.

In the one-dimensional specialization, this question has a particularly clean answer. Let F and G denote the distribution functions of the signal distribution P and the noisy distribution

Q. The exact distributional denoiser is the monotone transport map

$$T_\infty(y) = F^{-1} \circ G(y), \qquad (T_\infty)_\sharp Q = P.$$

This map is the one-dimensional Brenier map. It is the map we would like to use, but it is written in terms of F, the distribution function of the unobserved signal distribution.

The striking observation in Liang [2025b] is that, in the small-noise regime, this transport map can nevertheless be expanded using information from the noisy distribution alone. Writing $\alpha = \frac{\sigma^2}{2}$, Liang characterizes the Brenier map

$$T_\infty(y) = y + \sum_{k \geq 1} \frac{\alpha^k}{k!} h_k(y).$$

The coefficients, h_k, can be computed recursively from higher-order score functions of the noisy density,

$$\frac{q^{(m)}(y)}{q(y)},$$

with the recursion governed by partial Bell polynomials. Accordingly, a K-th order approximation can be derived via truncation

$$T_K(y) = y + \sum_{k=1}^{K} \frac{\alpha^k}{k!} h_k(y).$$

This is precisely the local-to-identity formulation we conjectured. As K increases, the denoisers successively approximate the monotone transport map; at the infinite-order endpoint, the hierarchy recovers T_∞.

First- and second-order distributional denoisers. The all-orders hierarchy is specialized to the one-dimensional Gaussian setting. The first two corrections, however, persist more broadly. In Liang [2025a], the author studies the multidimensional additive model with known noise level σ, allowing the signal and noise distributions to vary within smoothness

and moment classes.

In this setting, the Monge–Ampére criterion will separate the local transport correction motivated above from the classical Bayes action. For convenience, we let q denote the density of the noisy distribution, and write $s(y) = \nabla \log q(y)$. The author proposes the following first- and second-order distributional denoisers

$$\mathrm{T}_1(y) = y + \frac{\sigma^2}{2} s(y),$$

$$\mathrm{T}_2(y) = y + \frac{\sigma^2}{2} s(y) - \frac{\sigma^4}{8} \nabla \left\{ \frac{1}{2} \|s(y)\|^2 + \nabla \cdot s(y) \right\}.$$

The first map is the midpoint between the identity and the Tweedie denoiser. The second adds the next local correction, built from derivatives of the noisy score.

Liang evaluates these maps through the Monge–Ampère discrepancy. Under some assumptions we neglect in this exposition, a series of bounds create a striking contrast.

$$d_{\mathrm{MA}}\big(\mathrm{T}_1 \sharp P_Y, P_X \big) \leq C_\Omega \sigma^4.$$

$$d_{\mathrm{MA}}\big(\mathrm{T}_2 \sharp P_Y, P_X \big) \leq C_\Omega \sigma^6.$$

By comparison, in the Gaussian case the Bayes rule

$$\mathrm{T}^\star(y) = y + \sigma^2 s(y)$$

generally leaves a second-order Monge–Ampère discrepancy.

$$d_{\mathrm{MA}}\big(\mathrm{T}^\star \sharp P_Y, P_X \big) \geq c_\Omega \sigma^2.$$

The Bayes rule remains the optimal pointwise denoiser (see Chapter 2), while T_1 and T_2

are the low-order corrections selected by the distributional transport criterion.

Implications for diffuse–then–denoise. The preceding maps are local. Their relevance to diffusion comes from the fact that a diffusion sampler is built by composing local denoising steps. Consider again one Ornstein–Uhlenbeck step,

$$Y = (1-\eta)X_t + \sqrt{2\eta}\, Z.$$

The additive-noise variance in this step is 2η. If $q_{t+\eta}$ is the density of the noisy distribution and

$$s_{t+\eta}(y) = \nabla \log q_{t+\eta}(y),$$

then Liang's first distributional correction gives, for the scaled previous state,

$$y + \eta s_{t+\eta}(y).$$

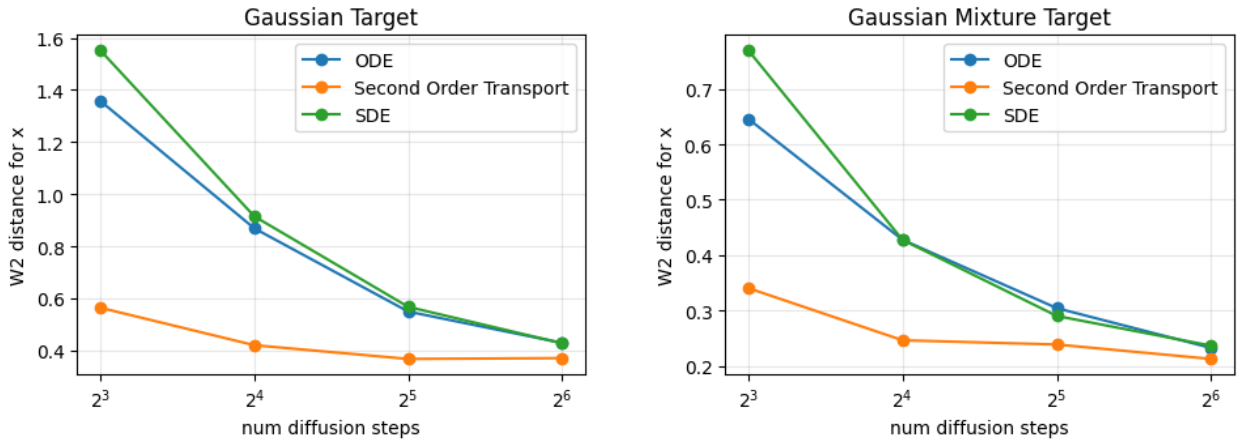
Undoing the linear scaling gives the local reverse map

$$D^{(1)}_{t,\eta}(y) = \frac{1}{1-\eta}\left\{y + \eta s_{t+\eta}(y)\right\}.$$

The second-order correction gives

$$D^{(2)}_{t,\eta}(y) = \frac{1}{1-\eta}\left[y + \eta s_{t+\eta}(y) - \frac{\eta^2}{2}\nabla\left\{\frac{1}{2}\|s_{t+\eta}(y)\|^2 + \nabla\cdot s_{t+\eta}(y)\right\}\right].$$

This suggests a new schemata for diffuse–then–denoise sampling. One first runs the forward noising chain to produce terminal samples at a large time. These samples initialize the backward chain. The reverse path is then obtained by applying the local maps $D^{(1)}_{t,\eta}$ or $D^{(2)}_{t,\eta}$ at successive noise levels. In this way, the usual reverse diffusion step is replaced by a denoising map chosen through the distributional transport criterion.



(a) Target $\mu = \mathcal{N}(0, 4)$. *(b) Target $\mu = \frac{1}{2}\mathcal{N}(-3, 2) + \frac{1}{2}\mathcal{N}(3, 2)$*

Figure 4.1: Schematic comparison of local reverse maps in a diffuse–then–denoise procedure. The second-order map adds a local correction involving derivatives of the score. We compare with the standard ODE/SDE reverse schemes across a range noise steps.

This is a fresh perspective on diffusion sampling motivated via a new optimality criterion. The forward chain creates a sequence of small-noise denoising problems. The Monge–Ampère criterion specifies how each local map should move probability mass. The resulting reverse path is therefore not only a composition of pointwise denoisers, but a composition of near-identity transport corrections that target toward the optimal transport from noise to signal.

4.3 From Local Transport to Generation

Diffusion models frame sampling as the problem of moving a simple reference distribution to a target distribution. Let ν be a tractable reference distribution and let μ be the distribution we wish to generate. A sampler is a mechanism for turning draws from ν into draws from μ. In principle, this is a global problem: the reference and target distributions may be far apart, and the transformation carrying one to the other may be highly complex.

Diffusion models make this problem tractable by decomposing it. Rather than construct a single map from ν to μ, the forward process builds a path of intermediate distributions,

$$\mu = \mu_0 \longrightarrow \mu_{t_1} \longrightarrow \cdots \longrightarrow \mu_{t_K} \approx \nu.$$

The reverse sampler then moves along this path in the opposite direction,

$$\nu \approx \mu_{t_K} \longrightarrow \mu_{t_{K-1}} \longrightarrow \cdots \longrightarrow \mu_0 = \mu.$$

The global sampling problem is replaced by a sequence of local denoising problems. Each reverse step moves between neighboring distributions, and each local map is close to the identity. This is the basic simplification: generation is organized as a composition of small transport corrections.

A parallel perspective appears in recent work on parabolic Monge–Ampère dynamics [Deb and Liang, 2026]. There, the aim is closer to the global optimal transport problem: construct a map that moves a reference distribution to a target distribution. The useful analogy is not in the details of the construction, but in the form of the strategy. Instead of solving for the full transport map at once, the method refines the map through an iterative procedure. A difficult global transformation is approached by a sequence of simpler updates.

This places the preceding discussion in a broader generative picture. Moving a simple reference distribution to a complex target distribution is, in general, a difficult measure transport problem. Modern generative methods often make this problem tractable by replacing one global transformation with a sequence of simpler local ones. This is the organizing idea in diffusion models, and it also appears in newer transport-based perspectives such as Deb and Liang [2026]. The complexity of the target distribution is handled gradually, through composition and iterative refinement, rather than through a single map learned all at once. This decomposition is one reason such methods can be applied to high-dimensional, multimodal, and non-log-concave distributions.

4.4 From Representation to Simulation

The denoising chapters treat the sampler as the object of study. We have asked how a generative model represents an observed distribution, how the reverse path behaves, and

how each local denoising step may be understood as measure transport.

The next part of the thesis uses the sampler differently. Once a distribution can be represented operationally, it can also be used computationally. Expectations, conditional expectations, and moment restrictions can be approximated by simulation from the fitted distribution. A conditional sampler is especially useful: by fixing the conditioning variables and sampling the remaining randomness, it creates a controlled synthetic environment inside the learned observational distribution.

This is the role of generative modeling in the next chapter. The target is no longer the data distribution itself, but a structural function characterized by conditional moments,

$$\mathbb{E}[\psi(O, \theta_0) \mid S] = 0.$$

A learned conditional sampler makes these restrictions operational by allowing the statistician to simulate from the fitted conditional distribution at selected support points.

The movement is therefore from representation to use. In the first part, the sampler is studied as a representation of an observed environment. In the next part, it becomes a statistical instrument for evaluating another inferential target.

CHAPTER 5

SIMULATING OBSERVED STRUCTURE: MOMENT RESTRICTIONS AND GENERATIVE INSTRUMENTAL VARIABLES

5.1 Introduction

Many econometric models are built from restrictions on conditional expectations. Economic assumptions often say that, after conditioning on the right variables, a structural error has mean zero. Instrument validity, rational-expectations restrictions, Euler equations, and many treatment or selection models all take this form. These assumptions do not require specifying the full distribution of the data, but they do generate moment restrictions that can identify structural objects. The generalized method of moments provided a general framework for estimation from unconditional moments [Hansen, 1982], while nonlinear rational-expectations applications emphasized how such moments often originate from conditional restrictions [Hansen and Singleton, 1982]. A large literature then studied how to exploit conditional restrictions more directly, through optimal instruments, efficiency theory, and sieve or nonparametric estimation [Sargan, 1958, Chamberlain, 1987, Newey, 1993, Ai and Chen, 2003, Chen and Pouzo, 2012].

This chapter explores how generative models, and in particular conditional samplers, can be used to estimate models defined by conditional moment restrictions. We focus on instrumental variables as a canonical and self-contained setting. We observe i.i.d. data

$$(Y_i, X_i, Z_i, W_i),$$

where Y_i is an outcome, X_i is an endogenous regressor, Z_i is an instrument, and W_i denotes

observed covariates. The object of interest is a structural function g_0 satisfying

$$\mathbb{E}[Y - g_0(X, W) \mid Z, W] = 0.$$

In linear settings, this restriction leads to familiar finite-dimensional procedures such as two-stage least squares. Once g_0 is allowed to be nonlinear or heterogeneous, however, estimation becomes substantially more difficult. The restriction is conditional, so using it directly requires evaluating conditional expectations indexed by instruments and covariates. In nonparametric IV, this gives rise to an ill-posed inverse problem [Newey and Powell, 2003, Hall and Horowitz, 2005, Darolles et al., 2011]. Existing approaches therefore either regularize the inverse problem directly or approximate the conditional restriction with a collection of unconditional moments.

Our approach keeps the conditional restriction as the target object but changes how the conditional expectations are computed. We learn a conditional generator for the reduced-form law of $(Y, X) \mid Z, W$, use it to simulate from the relevant conditional distributions, and estimate the structural function by minimizing a simulation-based conditional moment criterion. The generative model provides a Monte Carlo integration device for evaluating the conditional expectations that appear in the IV restriction.

This reframes the first stage as a conditional sampling problem and the second stage as a conditional moment problem. The procedure remains close to the original identifying restriction while allowing both the structural function and the reduced-form first-stage law to be flexible. This is precisely where generative models are useful: diffusion models, for example, are designed to approximate complex non-Gaussian, multimodal, and high-dimensional distributions [Liang et al., 2026]. The resulting estimator is also computationally tractable. Independent conditional batches yield unbiased Monte Carlo gradients for the squared conditional moment target, making the second stage amenable to standard gradient-based optimization.

The chapter develops theory and evidence in three steps. First, we prove a consistency result for a compact, fixed-dimensional series class. In this setting, Wasserstein convergence of the learned conditional law controls the residual-class integration error that links the first-stage sampler to the second-stage criterion. The result is intentionally modest, but it makes precise the first-stage property needed by the IV estimator.

Second, we evaluate the method in synthetic nonlinear IV designs, including scalar, multivariate, and multimodal first-stage settings with observed side information. Third, we revisit the returns-to-schooling design based on college proximity [Card, 1995, 2001]. This application is natural for our framework: the first stage may be heterogeneous across observed covariates, and the return to schooling need not be well summarized by a single constant marginal effect.

5.2 Instrumental Variables and Conditional Moment Restrictions

We begin by stating our data generating model and defining the statistical object we aim to recover. Suppose we observe i.i.d. data

$$(Y_i, X_i, Z_i, W_i),$$

where $Y_i \in \mathbb{R}$ is the outcome, $X_i \in \mathcal{X}$ is an endogenous regressor, $Z_i \in \mathcal{Z}$ is an excluded instrument, and $W_i \in \mathcal{W}$ is a vector of observed covariates. The structural equation is

$$Y = g_0(X, W) + \varepsilon, \tag{5.2.1}$$

where g_0 is the structural function of interest and ε is an unobserved disturbance. Endogeneity arises because X may be statistically dependent on ε.

The instrumental variables framework imposes three conditions. First, Z is relevant for

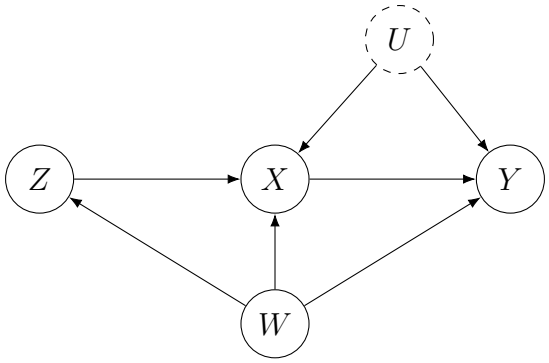


Figure 5.1: Causal diagram for the instrumental variables model with observed covariates.

X conditional on W, in the sense that the conditional law of X varies with Z:

$$P(X \mid Z, W) \neq P(X \mid W). \tag{5.2.2}$$

Second, Z satisfies exclusion: conditional on (X, W), it affects Y only through (5.2.1). Third, Z is exogenous:

$$\mathbb{E}[\varepsilon \mid Z, W] = 0. \tag{5.2.3}$$

Under (5.2.1) and (5.2.3), the structural function satisfies the conditional moment restriction

$$\mathbb{E}[Y - g_0(X, W) \mid Z, W] = 0. \tag{5.2.4}$$

This is the nonlinear IV model studied throughout the chapter. It accommodates nonlinear and heterogeneous structural effects and places the problem in the general class of conditional moment restriction models; see Chamberlain [1987], Ai and Chen [2003].

Figure 5.1 summarizes the underlying causal structure. The latent variable U represents unobserved heterogeneity that may jointly affect X and Y. The disturbance ε collects the component of that heterogeneity that enters the outcome equation outside the channel through X, together with any additional idiosyncratic noise.

5.2.1 Conditional moments as an inverse problem

The restriction (5.2.4) is infinite-dimensional: it must hold almost surely over the support of (Z, W). Let

$$(Tg)(z, w) := \mathbb{E}[g(X, W) \mid Z = z, W = w]$$

denote the conditional expectation operator, and define

$$m_0(z, w) := \mathbb{E}[Y \mid Z = z, W = w].$$

Then (5.2.4) is equivalent to

$$Tg_0 = m_0. \tag{5.2.5}$$

Nonlinear IV estimation is therefore an inverse problem: one seeks to recover the structural function g_0 from its image under the conditional expectation operator T; see Newey and Powell [2003], Hall and Horowitz [2005], Darolles et al. [2011].

This inverse problem is difficult because T is a smoothing operator. Changes in g are averaged over the conditional law of $X \mid Z, W$, so small errors in the estimated conditional mean m_0, or in the estimated operator itself, can be amplified when solving for g_0. The severity of this amplification is governed by the ill-posedness of T, often described through the decay of its singular values. Direct NPIV estimators therefore require regularization, and their performance depends on the conditioning of the operator together with tuning choices such as bandwidths, penalties, and sieve dimensions.

The same operator perspective separates identification from stability. Identification requires that T be injective on the relevant function class. A standard sufficient condition is completeness: if

$$\mathbb{E}[h(X, W) \mid Z, W] = 0 \quad \text{a.s.} \tag{5.2.6}$$

implies

$$h(X, W) = 0 \quad \text{a.s.},$$

then (5.2.4) uniquely determines g_0. Completeness requires the instrument to induce sufficiently rich variation in the conditional law of $X \mid Z, W$; see Newey and Powell [2003], Freyberger [2017]. But uniqueness alone does not guarantee stable estimation. Even when $Tg = Tg_0$ implies $g = g_0$, small perturbations in Tg may correspond to large perturbations in g. Thus nonlinear IV requires both identifying variation and a method for stabilizing the inverse problem.

5.2.2 From conditional to unconditional moments

A common way to avoid direct estimation of (5.2.4) is to work with unconditional orthogonality conditions. For any square-integrable test function $\phi(Z, W)$,

$$\mathbb{E}[(Y - g_0(X, W))\phi(Z, W)] = 0. \tag{5.2.7}$$

Thus the conditional moment restriction induces an infinite family of unconditional moments indexed by instrument functions ϕ.

Sieve IV and related finite-dimensional procedures make this representation computational by choosing a finite collection of instruments

$$Q_L(Z, W) = (q_1(Z, W), \ldots, q_L(Z, W))^\top$$

and imposing

$$\mathbb{E}[Q_L(Z, W)\{Y - g(X, W)\}] = 0.$$

This replaces the full conditional restriction by its projection onto the span of the chosen instrument basis. The resulting estimator can be stable and computationally simple, but its performance depends on whether the selected basis captures the components of the conditional moment restriction that identify the structural function. If the basis is too small, relevant conditional variation is omitted; if it is too rich, the sample moment matrices can

become ill-conditioned and the estimator may inherit the instability of the underlying inverse problem.

This tension motivates the approach below. Rather than choosing a finite instrument basis or directly inverting an estimated operator, we learn a conditional sampler and use it to evaluate the conditional moments appearing in (5.2.4) by simulation.

5.3 Generative IV

The preceding section described two standard routes to tractability: regularize the inverse problem directly, or replace the conditional restriction by finitely many unconditional moments. We take a different route. The conditional restriction remains the target; a learned conditional sampler makes its moments directly computable by simulation. Rather than reformulating the restriction, we optimize a simulated loss built from the conditional moments themselves.

This section first motivates a conditional-moment loss that targets the restriction directly. It then introduces a tractable finite-sample analogue and a gradient-based optimization method for solving it. Finally, it isolates the approximation error induced by replacing the true conditional law with the fitted sampler and discusses its implications for first-stage learning.

5.3.1 From moment restriction to loss criterion

The IV conditional moment restriction (5.2.4) suggests a direct target: choose a structural function whose conditional moment vanishes. We impose this restriction over a parametrized structural class

$$\mathcal{G}_\Theta = \{g_\theta : \theta \in \Theta\},$$

which may include linear, polynomial, spline, or neural-network specifications. Let $s = (z, w)$ denote a realization of $S = (Z, W)$, and write

$$P_s := P(\cdot \mid S = s)$$

for the conditional law of $(Y, X) \mid S = s$. Define the residual

$$\psi_\theta(y, x, w) := g_\theta(x, w) - y.$$

The conditional moment at s is

$$m_s(\theta) := \int \psi_\theta(y, x, w)\, dP_s(y, x) = \mathbb{E}[g_\theta(X, w) - Y \mid S = s].$$

The IV restriction is $m_S(\theta_0) = 0$ almost surely. We therefore define the population conditional moment criterion

$$L(\theta) := \mathbb{E}[m_S(\theta)^2]. \tag{5.3.1}$$

Under correct specification, any parameter satisfying the IV restriction attains the minimum value 0. More generally, the population target is the set of minimizers of $L(\theta)$ over Θ.

5.3.2 Finite-sample criteria and gradient-based optimization

The population criterion (5.3.1) is infeasible because the conditional moments $m_s(\theta)$ are unknown. We approximate it on a finite collection of conditioning values, or support points, $s_i = (z_i, w_i)$, $i = 1, \ldots, n$. In applications these support points may be observed values of $S_i = (Z_i, W_i)$, or a subsample thereof.

If the conditional laws P_{s_i} were known, the oracle support-point criterion would be

$$L_n^P(\theta) := \frac{1}{n} \sum_{i=1}^{n} m_i^P(\theta)^2, \qquad m_i^P(\theta) := m_{s_i}(\theta). \tag{5.3.2}$$

When the support points are sampled from the marginal law of S, L_n^P is the empirical analogue of the population loss $L(\theta)$.

In practice, the conditional laws P_{s_i} are unknown. Given a fitted conditional sampler, write

$$\widehat{P}_i := \widehat{P}(\cdot \mid S = s_i)$$

and define

$$m_i^{\widehat{P}}(\theta) := \int \{g_\theta(x, w_i) - y\}\, d\widehat{P}_i(y, x).$$

The generator-induced support-point criterion is

$$L_n^{\widehat{P}}(\theta) := \frac{1}{n} \sum_{i=1}^{n} m_i^{\widehat{P}}(\theta)^2. \tag{5.3.3}$$

This is the integrated target of the feasible estimator. Conditional Monte Carlo is used to approximate its moments and construct gradients centered on this target.

Direct plugin simulation. Conditional on $\widehat{P}$, draw

$$(Y_{ij}, X_{ij}) \sim \widehat{P}_i, \qquad j = 1, \ldots, m,$$

and set

$$\psi_{ij}(\theta) := g_\theta(X_{ij}, w_i) - Y_{ij}, \qquad \bar{\psi}_i(\theta) := \frac{1}{m} \sum_{j=1}^{m} \psi_{ij}(\theta).$$

The direct simulated analogue of (5.3.3) is

$$\widehat{L}_{\text{plug}}^{\widehat{P}}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \bar{\psi}_i(\theta)^2. \tag{5.3.4}$$

Although natural, this plugin criterion has a gradient that is not centered on $\nabla_\theta L_n^{\widehat{P}}(\theta)$. Conditional on $s_1, \ldots, s_n$ and $\widehat{P}$,

$$\mathbb{E}\Big[\nabla_\theta \widehat{L}^{\widehat{P}}_{\text{plug}}(\theta) \mid s_1, \dots, s_n, \widehat{P}\Big] = \nabla_\theta L_n^{\widehat{P}}(\theta) + \frac{2}{nm} \sum_{i=1}^{n} \operatorname{Cov}_{\widehat{P}}(g_\theta(X, w_i) - Y,\, \nabla_\theta g_\theta(X, w_i) \mid S = s_i)\,. \tag{5.3.5}$$

The extra term is a diagonal Monte Carlo bias.

Two-batch correction. To remove this diagonal term, draw two independent conditional batches from $\widehat{P}_i$:

$$\{(Y_{ij}, X_{ij})\}_{j=1}^m \quad \text{and} \quad \{(Y'_{ij}, X'_{ij})\}_{j=1}^m.$$

Let primed quantities denote averages over the second batch. Define

$$\dot{\psi}_{ij}(\theta) := \nabla_\theta g_\theta(X_{ij}, w_i), \qquad \bar{\dot{\psi}}_i(\theta) := \frac{1}{m} \sum_{j=1}^{m} \dot{\psi}_{ij}(\theta),$$

and define $\dot{\psi}'_{ij}(\theta)$ and $\bar{\dot{\psi}}'_i(\theta)$ analogously.

The two-batch gradient estimator is

$$\widehat{G}_U^{\widehat{P}}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \Big\{ \bar{\dot{\psi}}_i(\theta) \bar{\psi}'_i(\theta) + \bar{\psi}_i(\theta) \bar{\dot{\psi}}'_i(\theta) \Big\}. \tag{5.3.6}$$

Independence of the two batches gives

$$\mathbb{E}[\widehat{G}_U^{\widehat{P}}(\theta) \mid s_1, \dots, s_n, \widehat{P}] = \nabla_\theta L_n^{\widehat{P}}(\theta). \tag{5.3.7}$$

Equivalently, $\widehat{G}_U^{\widehat{P}}$ is the gradient of the two-batch criterion

$$\widehat{L}_U^{\widehat{P}}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \bar{\psi}_i(\theta) \bar{\psi}'_i(\theta), \tag{5.3.8}$$

which satisfies

$$\mathbb{E}[\widehat{L}_U^{\widehat{P}}(\theta) \mid s_1, \ldots, s_n, \widehat{P}] = L_n^{\widehat{P}}(\theta). \tag{5.3.9}$$

The two-batch objective need not be nonnegative in finite samples. Its role is to provide a diagonal-corrected Monte Carlo gradient for the feasible squared conditional moment target. The identities in (5.3.5)–(5.3.9) are proved in Appendix B.3.1.

5.3.3 First-stage approximation error

The fitted law $\widehat{P}(\cdot \mid S = s)$ can be any conditional sampler for the reduced-form law of $(Y, X) \mid S = s$. In our implementation, it is a conditional score-based generative model [Dharmakeerthi et al., 2025]. The sampler is used as a conditional integration device: at each support point $s_i = (z_i, w_i)$, it supplies draws from the fitted law of $(Y, X) \mid S = s_i$ for evaluating the moments entering $L_n^{\widehat{P}}$.

The preceding subsection shows how conditional Monte Carlo provides gradients centered on the generator-induced target $L_n^{\widehat{P}}$. It remains to relate this target to the oracle criterion L_n^P. The following decomposition isolates the error from replacing the true conditional law by the fitted sampler.

Proposition 5.3.1. *Fix support points $s_1, \ldots, s_n$. Write*

$$P_i := P(\cdot \mid S = s_i), \qquad \widehat{P}_i := \widehat{P}(\cdot \mid S = s_i),$$

and define

$$\delta_i(\theta) := m_i^{\widehat{P}}(\theta) - m_i^P(\theta), \qquad D_n(\theta)^2 := \frac{1}{n}\sum_{i=1}^n \delta_i(\theta)^2.$$

Then

$$L_n^{\widehat{P}}(\theta) - L_n^P(\theta) = \frac{2}{n}\sum_{i=1}^n m_i^P(\theta)\delta_i(\theta) + D_n(\theta)^2, \tag{5.3.10}$$

and hence

$$|L_n^{\widehat{P}}(\theta) - L_n^P(\theta)| \le 2\{L_n^P(\theta)\}^{1/2} D_n(\theta) + D_n(\theta)^2. \tag{5.3.11}$$

Algorithm 5.3.1 Generative IV estimation

Require: Training data $\{(Y_i, X_i, S_i)\}_{i=1}^N$, and support points $s_1, \ldots, s_n$, Monte Carlo size m, structural class $\{g_\theta : \theta \in \Theta\}$, optimization steps T

Ensure: Structural estimate $g_{\widehat{\theta}}$

1: Fit a conditional sampler $\widehat{P}(\cdot \mid S = s)$ for the reduced-form law of $(Y, X) \mid S = s$
2: Initialize θ_0
3: **for** $t = 0, \ldots, T-1$ **do**
4: Select support-point batch $\mathcal{B}_t \subseteq \{1, \ldots, n\}$
5: **for** $i \in \mathcal{B}_t$ **do**
6: Draw two independent conditional batches from $\widehat{P}(\cdot \mid S = s_i)$

$$\{(Y_{ij}, X_{ij})\}_{j=1}^m, \qquad \{(Y'_{ij}, X'_{ij})\}_{j=1}^m.$$

7: Form residual averages $\bar{\psi}_i(\theta_t)$ and $\bar{\psi}'_i(\theta_t)$
8: **end for**
9: Update θ_t using the gradient of

$$\widehat{L}_{U,t}^{\widehat{P}}(\theta_t) := \frac{1}{|\mathcal{B}_t|} \sum_{i \in \mathcal{B}_t} \bar{\psi}_i(\theta_t) \bar{\psi}'_i(\theta_t).$$

10: **end for**
11: **return** $g_{\widehat{\theta}}$, where $\widehat{\theta} = \theta_T$

Proposition 5.3.1 shows that the first-stage error enters the second-stage criterion through the residual integration error

$$D_n(\theta)^2 = \frac{1}{n} \sum_{i=1}^n \left[\int \{g_\theta(x, w_i) - y\} \, d(\widehat{P}_i - P_i)(y, x) \right]^2.$$

Thus the sampler need not approximate every feature of the conditional law equally well. For the second-stage criterion, the relevant first-stage property is accurate integration of the residual class

$$\Psi_\Theta := \{(y, x, w) \mapsto g_\theta(x, w) - y : \theta \in \Theta\}.$$

Section 5.4 gives a consistency result for a bounded finite-dimensional series class in which this residual-class discrepancy is controlled by average conditional Wasserstein convergence of the sampler.

5.4 A Consistency Result

This section gives a benchmark consistency result for the generative-IV criterion. To isolate the role of first-stage approximation, we work with a compact, fixed-dimensional series class.

Consider

$$\mathcal{G}_C := \left\{ g_\theta(x,w) = p_K(x,w)^\top \theta : \theta \in \Theta_C \right\}, \qquad \Theta_C := \left\{ \theta \in \mathbb{R}^K : \|\theta\|_2 \le C \right\}, \tag{5.4.1}$$

where p_K is a fixed K-dimensional vector of linear, polynomial, spline, or other series terms.

Assumption 5.4.1 (Compact support)**.** There exist compact sets

$$\mathcal{X}_0 \subset \mathbb{R}^{d_x}, \qquad \mathcal{W}_0 \subset \mathbb{R}^{d_w},$$

such that $W \in \mathcal{W}_0$ almost surely. For each support point $s_i = (z_i, w_i)$, the conditional laws

$$P_i := P(\cdot \mid S = s_i), \qquad \widehat{P}_i := \widehat{P}(\cdot \mid S = s_i),$$

viewed as distributions over (Y, X), assign probability one to

$$\mathbb{R} \times \mathcal{X}_0.$$

Assumption 5.4.2 (Bounded basis)**.** The basis map

$$p_K : \mathcal{X}_0 \times \mathcal{W}_0 \to \mathbb{R}^K$$

is uniformly bounded and uniformly Lipschitz in x. That is, there exist constants $M_p < \infty$ and $L_p < \infty$ such that

$$\sup_{(x,w) \in \mathcal{X}_0 \times \mathcal{W}_0} \|p_K(x,w)\|_2 \le M_p, \tag{5.4.2}$$

and, for all $x, x' \in \mathcal{X}_0$ and all $w \in \mathcal{W}_0$,

$$\|p_K(x, w) - p_K(x', w)\|_2 \le L_p \|x - x'\|. \tag{5.4.3}$$

Assumption 5.4.1 is a technical support condition. Assumption 5.4.2 is satisfied by fixed-degree polynomial or spline bases on compact support.

Specializing the residual class from Section 5.3 to Θ_C, define

$$\Psi_C := \{\psi_\theta : \theta \in \Theta_C\}.$$

The corresponding residual-class discrepancy is

$$\Delta_{\Psi_C,n}(\widehat{P}, P) := \sup_{\theta \in \Theta_C} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[\int \psi_\theta(y, x, w_i)\, d(\widehat{P}_i - P_i)(y, x) \right]^2 \right\}^{1/2}. \tag{5.4.4}$$

Proposition 5.4.3 (Residual-class control for bounded series)**.** *Suppose Assumptions 5.4.1 and 5.4.2 hold, and suppose the conditional laws have finite first moments. Then, for every $\theta \in \Theta_C$, every $w \in \mathcal{W}_0$, every $x, x' \in \mathcal{X}_0$, and every $y, y' \in \mathbb{R}$,*

$$|\psi_\theta(y, x, w) - \psi_\theta(y', x', w)| \le L_C \|(y, x) - (y', x')\|, \qquad L_C := \sqrt{1 + C^2 L_p^2}.$$

Consequently,

$$\Delta_{\Psi_C,n}(\widehat{P}, P) \le L_C \left\{ \frac{1}{n} \sum_{i=1}^{n} W_1(\widehat{P}_i, P_i)^2 \right\}^{1/2}. \tag{5.4.5}$$

If the conditional laws have finite second moments, then

$$\Delta_{\Psi_C,n}(\widehat{P}, P) \le L_C \left\{ \frac{1}{n} \sum_{i=1}^{n} W_2(\widehat{P}_i, P_i)^2 \right\}^{1/2}. \tag{5.4.6}$$

Proposition 5.4.3 converts average conditional Wasserstein control into the residual-class control required by Proposition 5.3.1. Thus, for the bounded series class (5.4.1), average

conditional W_1-convergence is sufficient for the first-stage requirement used below; W_2-convergence is also sufficient under finite second moments. This condition can be viewed as the form of first-stage accuracy that a conditional generative model must deliver; verifying it for trained score-based samplers requires assumptions beyond the scope of this result.

Restrict the population loss $L(\theta)$ in (5.3.1) to Θ_C, and define

$$\Theta_0 := \arg\min_{\theta \in \Theta_C} L(\theta), \qquad L_0 := \inf_{\theta \in \Theta_C} L(\theta).$$

The estimator analyzed below minimizes the exact generator-induced support-point criterion $L_n^{\widehat{P}}$ in (5.3.3), restricted to Θ_C. Thus the result abstracts from finite Monte Carlo and optimization error, and isolates the effect of replacing P_i by $\widehat{P}_i$.

Theorem 5.4.4 (Generative IV: consistency on a bounded series class)**.** *Suppose $s_1, \ldots, s_n$ are i.i.d. draws from the marginal law of S. Suppose Assumptions 5.4.1 and 5.4.2 hold, and suppose*

$$\mathbb{E}[\alpha(S)^2] < \infty, \qquad \alpha(s) := \mathbb{E}[Y \mid S = s].$$

Suppose further that

$$\left\{ \frac{1}{n} \sum_{i=1}^{n} W_1(\widehat{P}_i, P_i)^2 \right\}^{1/2} = o_p(1). \tag{5.4.7}$$

Finally, assume the population criterion separates Θ_0: for every $\varepsilon > 0$,

$$\inf_{\{\theta \in \Theta_C : \operatorname{dist}(\theta, \Theta_0) \geq \varepsilon\}} \{L(\theta) - L_0\} > 0,$$

where

$$\operatorname{dist}(\theta, \Theta_0) := \inf_{\vartheta \in \Theta_0} \|\theta - \vartheta\|_2.$$

If

$$\widehat{\theta} \in \arg\min_{\theta \in \Theta_C} L_n^{\widehat{P}}(\theta)$$

is a measurable minimizer, then

$$\operatorname{dist}(\widehat{\theta}, \Theta_0) \to_p 0.$$

If $\Theta_0 = \{\theta_0\}$*, then* $\widehat{\theta} \to_p \theta_0$*.*

Remark 5.4.5 (Interpretation). The set Θ_0 is the population minimizer set over the restricted class Θ_C. Under correct specification and identification within Θ_C, it is the identified structural set; under misspecification, it is the pseudo-true set.

The first-stage condition (5.4.7) is imposed under the joint law of the training data and support points. With sample splitting, it is an out-of-sample approximation condition. Without sample splitting, it is an in-sample approximation condition and requires control of overfitting.

The proof is given in Appendix B.3.2. Extensions to growing sieves, neural structural functions, or finite Monte Carlo optimization require additional empirical-process and regularization arguments.

5.5 Synthetic Experiments

We evaluate the estimator in three nonlinear IV designs. The designs are chosen to separate three sources of difficulty: nonlinear structural recovery, multivariate treatment effects, and heterogeneous or multimodal first-stage distributions. In each design, endogeneity is generated by a latent variable U that enters both the treatment equation and the outcome equation. The instrument Z shifts the conditional distribution of X, while the exclusion restriction is maintained by construction.

Throughout, the outcome takes the form

$$Y = g_0(X, W) + \rho U + \sigma_\varepsilon \varepsilon, \tag{5.5.1}$$

where ε is independent of (Z, W, U). The object of interest is the structural function g_0.

Performance is therefore evaluated by structural recovery error, rather than prediction error for Y.

Design 1: scalar nonlinear IV. The first design is a scalar nonlinear benchmark. We draw

$$Z, U, V, \varepsilon \sim \mathcal{N}(0, 1)$$

independently and set

$$X = Z + 0.8U + 0.3V, \qquad Y = g_1(X) + 0.8U + 0.3\varepsilon, \tag{5.5.2}$$

where

$$g_1(x) = 0.8\sin(1.2x) + 0.4\tanh(1.5x) + 0.15x^2. \tag{5.5.3}$$

This design favors classical nonlinear IV methods: the treatment is scalar, the first stage is smooth, and the structural function is nonlinear but low-dimensional.

Design 2: multivariate nonlinear IV. The second design keeps the first stage smooth but makes the structural object multivariate. We draw

$$Z = (Z_1, Z_2)^\top \sim \mathcal{N}(0, I_2), \qquad U \sim \mathcal{N}(0, 1), \qquad V = (V_1, V_2)^\top \sim \mathcal{N}(0, 0.4^2 I_2),$$

independently, and define

$$X_1 = 0.9Z_1 + 0.4Z_2 + 0.8U + V_1, \qquad X_2 = -0.3Z_1 + 0.8Z_2 + 0.8U + V_2. \tag{5.5.4}$$

The outcome is

$$Y = g_2(X_1, X_2) + 0.8U + 0.3\varepsilon, \qquad g_2(x_1, x_2) = 1.2x_1 + 0.4x_2^2 + 0.6x_1x_2. \tag{5.5.5}$$

This design tests whether the estimator can recover a joint structural surface rather than a scalar response curve.

Design 3: side information and a multimodal first stage. The third design introduces observed side information and a more complex first-stage distribution. We draw

$$W = (W_1, W_2)^\top \sim \mathcal{N}(0, I_2), \qquad Z = (Z_1, Z_2)^\top \sim \mathcal{N}(0, I_2), \qquad U \sim \mathcal{N}(0, 1),$$

and generate X from a context-dependent mixture distribution whose component weights and instrument strength vary with W. The outcome is

$$Y = g_3(X, W) + 0.7U + 0.3\varepsilon, \tag{5.5.6}$$

where

$$g_3(x, w) = x_1 + 0.5x_2^2 + 0.6w_1x_1 - 0.4w_2x_2 + 0.25x_1x_2. \tag{5.5.7}$$

This design is closest to the intended use case for the proposed estimator: the structural function depends on both treatment and observed covariates, and the first-stage law is not well summarized by a simple conditional mean.

5.5.1 Experimental Setup

For each design and estimator, we run $R = 10$ independent replications. The training sample sizes for the results in Table 5.1 are $n = 5000$ in Designs 1–3. The conditional moment criterion is evaluated on 500 support points, with $m = 500$ conditional Monte Carlo draws per support point.

We compare the proposed estimator with four benchmarks: polynomial OLS, linear IV, sieve 2SLS, and kernel NPIV [Darolles et al., 2011]. Polynomial OLS is included to show the effect of ignoring endogeneity while allowing nonlinear functional form. Linear IV is included

as the standard correctly instrumented but structurally restrictive benchmark. Sieve 2SLS and kernel NPIV are included as nonlinear IV benchmarks.

The primary evaluation metric is structural recovery error. For a fitted estimator $\widehat{g}$, we compute

$$\mathcal{E}(\widehat{g}) = \frac{1}{|\mathcal{G}|} \sum_{(x,w)\in\mathcal{G}} \{\widehat{g}(x,w) - g_0(x,w)\}^2,$$

where $\mathcal{G}$ is a deterministic grid over the interior of the relevant support. In Design 1, $\mathcal{G}$ is a grid over x. In Design 2, it is a grid over (x_1, x_2). In Design 3, the error averages slice-level mean squared errors over representative covariate profiles and treatment-coordinate slices.

5.5.2 Results

Table 5.1 reports the Monte Carlo results. The results show three patterns. First, methods that ignore endogeneity or impose a linear structural function perform poorly once the true structural function is nonlinear. Second, classical nonlinear IV methods are competitive when the treatment is low-dimensional and the first stage is smooth. Third, the proposed estimator performs best in the design where the first-stage conditional law is heterogeneous and multimodal, which is the setting most directly targeted by the generative construction.

In Design 1, kernel NPIV and sieve 2SLS perform very well. This is expected: the treatment is scalar, the first stage is smooth, and the structural function is low-dimensional. The proposed estimator is competitive but does not dominate the best classical nonlinear IV benchmark. This is an important sanity check. The generative procedure is not designed to replace simpler methods in settings where a low-dimensional inverse problem is already well behaved.

In Design 2, sieve 2SLS performs best because the true structural function is close to the polynomial class used by the estimator. Kernel NPIV deteriorates relative to Design 1, reflecting the additional difficulty of multivariate smoothing and regularized operator inversion. The proposed estimator remains competitive and substantially improves on polynomial

Table 5.1: Monte Carlo structural recovery error across the three synthetic designs. Entries report mean error across $R = 10$ replications. Design 3 was too computationally demanding for kernel NPIV. See appendix to see how performance changes depending on number of samples, n.

Method	Design 1	Design 2	Design 3
Polynomial OLS	0.4154 (0.0365)	0.3588 (0.0184)	0.2127 (0.0044)
Linear IV	0.3635 (0.0276)	2.6775 (0.1588)	4.0860 (1.6471)
Sieve 2SLS ($b = 10$)	0.0224 (0.0139)	**0.0039 (0.0023)**	0.1246 (0.0080)
Sieve 2SLS ($b = 1$)	0.6878 (0.0243)	2.0138 (0.1879)	0.8252 (0.1293)
Kernel NPIV	0.0548 (0.0312)	0.3263 (0.0182)	–
Generative IV	**0.0159 (0.0145)**	0.0136 (0.0082)	**0.0313 (0.0243)**

OLS, linear IV, and kernel NPIV.

Design 3 is the main stress test. The structural function depends on treatment and observed side information, while the first-stage law is heterogeneous and multimodal. In this setting, the proposed estimator performs best by a wide margin. This is the comparative advantage emphasized by the method: the first stage is modeled as a conditional distribution, not reduced to a small number of basis projections or a conditional mean, and the learned law is used directly inside the conditional moment criterion.

Figures 5.2–5.4 show representative replications. They are intended to complement the table by showing the shape of the recovered structural functions. The figures confirm that the proposed estimator tracks the structural curve or surface well in the more complex designs, rather than merely improving the average error.

5.6 College-Proximity and Returns to Schooling

We apply the method to the canonical college-proximity design of Card [1995, 2001]. The outcome is log wage, the endogenous regressor is years of schooling, and the excluded instrument is proximity to a college during adolescence. The identifying idea is that local access to college shifts schooling decisions, conditional on observed background and labor-market controls, while not directly entering wages except through schooling.

This application is useful for two reasons. First, it is a familiar benchmark for IV meth-

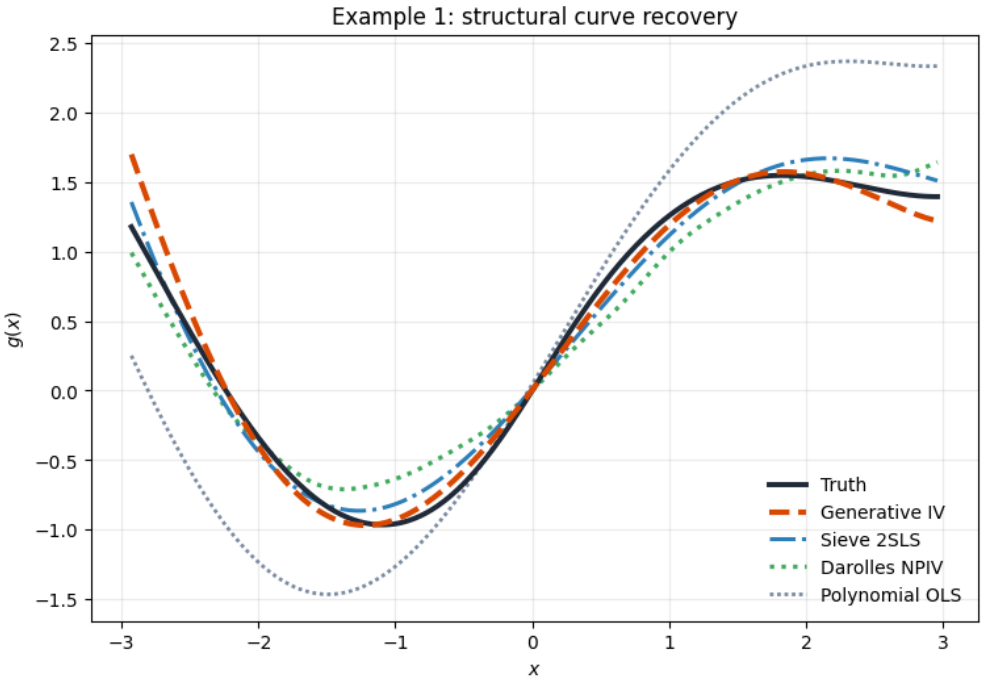


Figure 5.2: Representative replication for Design 1: scalar nonlinear IV.

ods. The linear specification provides a well-understood reference point: compare OLS with 2SLS and ask how much the return to schooling changes once schooling is instrumented. Second, the economic object need not be exhausted by a single constant return. The college-proximity instrument affects schooling on a particular margin, and the return for individuals shifted by that margin may differ from the average return in the population. This interpretation is consistent with the local-IV perspective of Imbens and Angrist [1994], the marginal-treatment-effect framework of Heckman and Vytlacil [2005], and the education literature emphasizing heterogeneous and marginal returns [Card, 2001, Carneiro et al., 2011].

5.6.1 Empirical Specification

Let Y denote log wage, X years of schooling, Z the college-proximity instrument, and W a vector of observed controls. We estimate

$$Y = g_0(X, W) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid Z, W] = 0.$$

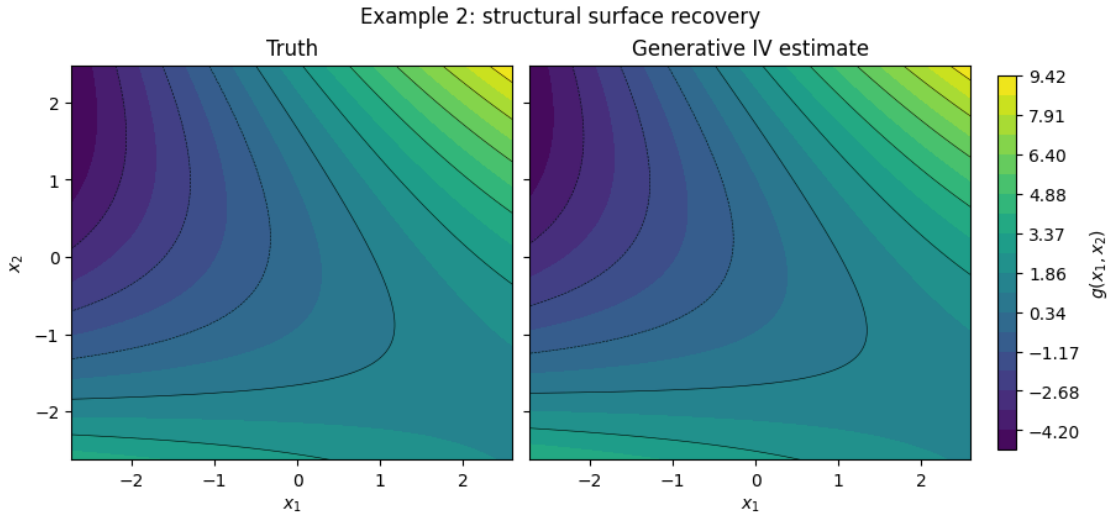


Figure 5.3: Representative replication for Design 2: multivariate nonlinear IV.

The control vector includes labor-market experience, location and indicators for other individual characteristics.

The empirical analysis has three components. We first report conventional linear benchmarks: OLS, 2SLS, and a linear version of the generative IV estimator. The last benchmark is important because it isolates the role of the sampler-based conditional moment procedure from the role of nonlinear structural flexibility. We then estimate the nonlinear generative IV model. The first-stage generator is fitted to the reduced-form conditional law of $(Y, X) \mid Z, W$, and the second stage uses the generated conditional draws to impose the conditional moment restriction.

5.6.2 Empirical Findings

Table 5.2 reports the linear estimates. OLS gives a return of about 7.4 percent, while linear 2SLS gives a larger estimate of about 13.2 percent. The linear generative IV estimate is very close to linear 2SLS. This is a useful validation exercise: when the structural class is restricted to be linear, the sampler-based conditional moment estimator recovers the familiar IV benchmark.

Figure 5.5 summarizes the first-stage content of the instrument. College proximity shifts

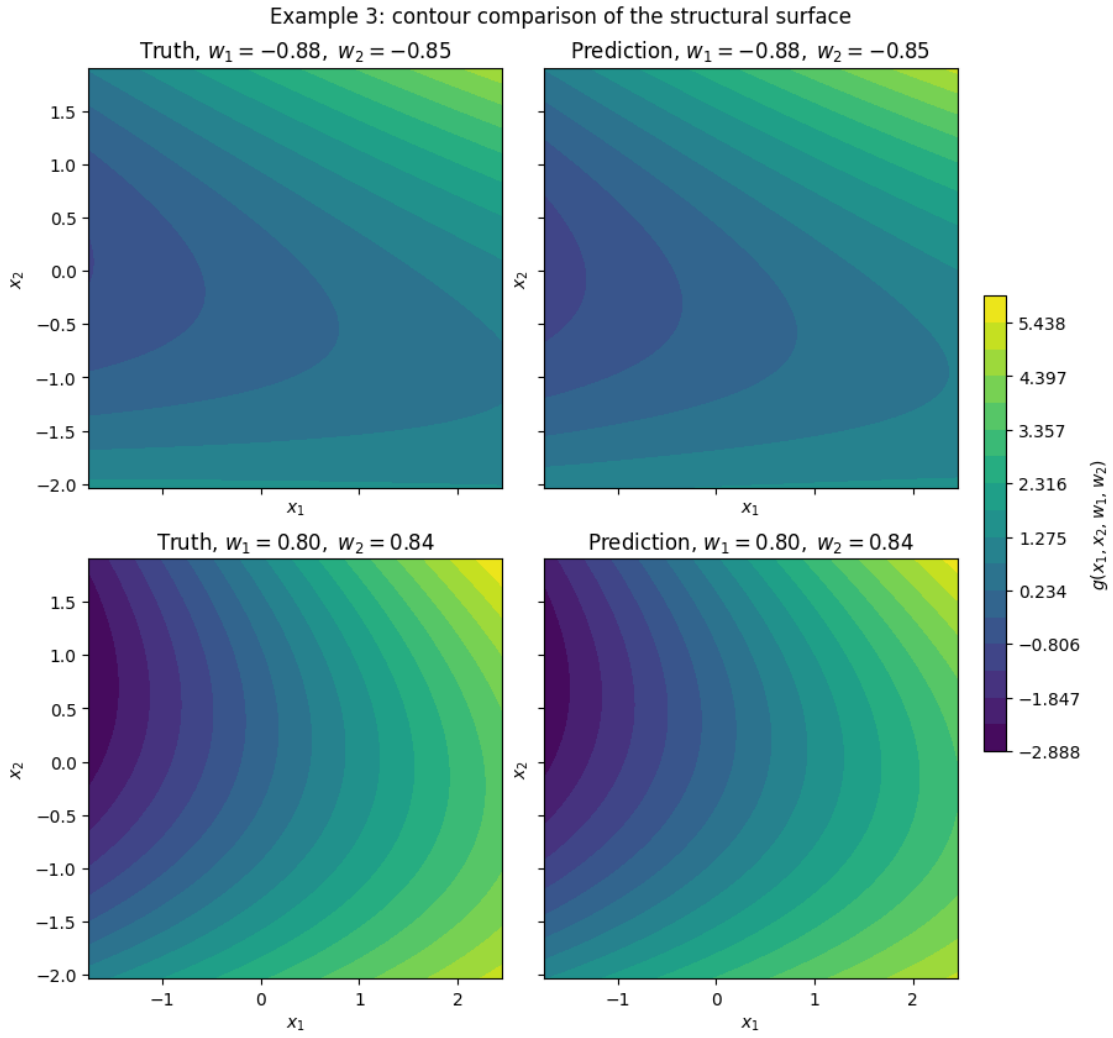


Figure 5.4: Representative replication for Design 3: side information and a heterogeneous multimodal first stage.

schooling, but the fitted conditional sampler suggests that this shift is not homogeneous across observed environments. This is economically important. If the instrument changes schooling more strongly for some covariate profiles than others, then a linear IV coefficient necessarily aggregates returns across a nonuniform margin of schooling choice. The generative first stage is useful because it represents this conditional distributional shift rather than reducing the first stage to a single conditional mean.

Figure 5.6 reports the main structural estimates. The left panel compares the schooling–earnings profiles implied by OLS, linear IV, linear generative IV, and nonlinear generative

IV. The right panel reports the corresponding marginal returns. The main empirical finding is that the nonlinear IV fit does not simply reproduce a constant return with a different level. Instead, it implies meaningful variation in the marginal return to schooling over the schooling distribution.

This pattern is consistent with several economic mechanisms emphasized in the returns-to-education literature. Years of schooling may have different labor-market value depending on whether they move individuals across completion or credential thresholds. A year that completes high school, begins college, completes college, or moves a worker toward post-graduate training may have a different return from a year that does not change the worker's credential or occupational opportunity set. Such nonlinearities are closely related to the literature on diploma or "sheepskin" effects [Hungerford and Solon, 1987, Belman and Heywood, 1991, Jaeger and Page, 1996] and to evidence that returns to postsecondary education vary across margins and institutional settings [Kane and Rouse, 1995, Oreopoulos and Petronijevic, 2013].

Figures 5.7 and 5.8 examine heterogeneity in the fitted nonlinear structural function. These figures should not be interpreted as separate subgroup-specific LATEs. Rather, they report model-implied heterogeneity in the structural function $g_0(x, w)$ under the maintained conditional moment restriction. The estimates suggest that the marginal return to schooling varies with both schooling and labor-market experience, with the largest departures from the constant-return benchmark occurring at low and high schooling levels. This is economically plausible: at low schooling levels, additional schooling may move individuals toward baseline labor-market credentials; at high schooling levels, additional schooling may open access to higher-skill occupations or more specialized career tracks. The heatmap makes this point visually by showing where the fitted derivative $\partial g(x, w)/\partial x$ is largest over the schooling-experience space.

Taken together, the application illustrates the practical value of our method. The linear generative IV estimate reproduces the conventional IV benchmark, while the nonlinear

Table 5.2: Linear benchmark estimates for the returns-to-schooling application.

Estimator	Coefficient
OLS	0.0740
Linear 2SLS	0.1323
Linear generative CMR	0.1301

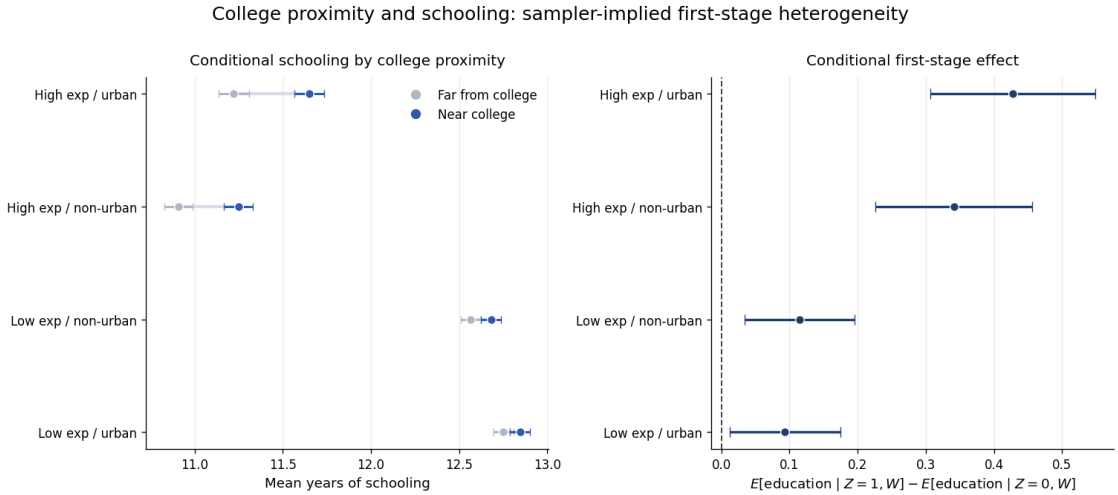


Figure 5.5: First-stage relevance in the college-proximity design. The left panel reports the average schooling shift associated with college proximity. The right panel uses the fitted conditional sampler to show how the implied first-stage shift varies across representative covariate profiles.

specification reveals structure that the linear approach suppresses. The contribution is not merely to allow a flexible function g_0. Our estimator combines structural flexibility with a flexible conditional first stage, then uses that first stage directly inside the conditional moment criterion. In the Card design, this yields an IV estimate that remains anchored to a classical source of identifying variation while allowing the return to schooling to vary across economically meaningful margins.

5.7 Conclusion

This chapter develops a generative approach to nonlinear instrumental variables estimation. The starting point is the conditional moment restriction itself. Instead of reducing that restriction to a fixed set of unconditional moments, the method learns a conditional sampler

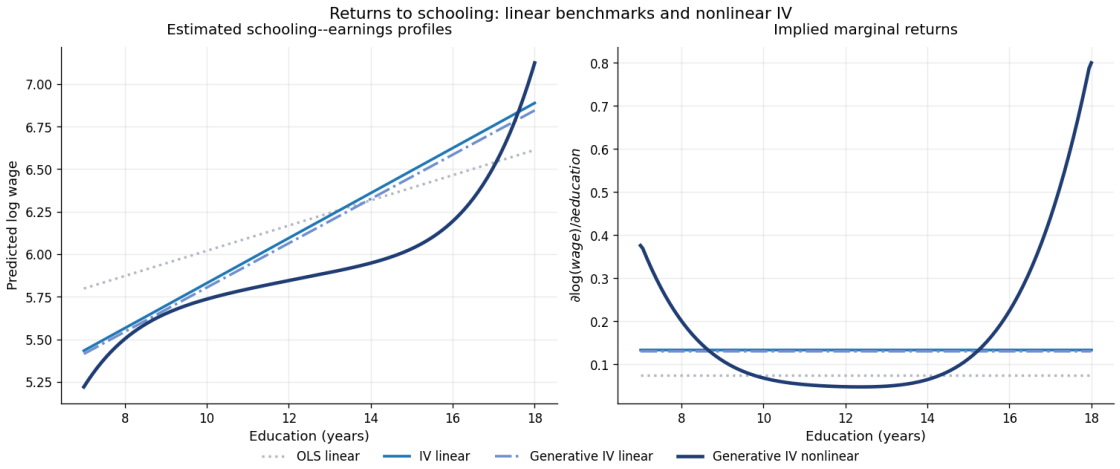


Figure 5.6: Structural schooling–earnings profiles and implied marginal returns. The left panel compares OLS, linear IV, linear generative IV, and nonlinear generative IV schooling curves. The right panel reports the corresponding marginal returns.

for the reduced-form law of $(Y, X) \mid Z, W$ and uses simulated conditional draws to evaluate the IV moment criterion.

The estimator separates the reduced-form and structural tasks. The first stage is a conditional distribution-learning problem; the second stage is a conditional moment problem. This separation allows the first-stage law to be non-Gaussian, multimodal, or high-dimensional while keeping the structural target anchored to the original IV restriction.

The theory shows that the first-stage requirement is residual-class integration. The learned sampler need not approximate the entire conditional law uniformly in every respect; it must approximate the conditional expectations of $g_\theta(X, W) - Y$ over the structural class. For bounded finite-dimensional series classes, Wasserstein convergence of the conditional sampler is sufficient for this condition and yields consistency of the generator-induced estimator.

The numerical evidence supports this view. Classical nonlinear IV methods remain competitive in simple low-dimensional designs, while the generative approach is most useful when the first-stage distribution is heterogeneous or multimodal. In the schooling application, the method reproduces the linear IV benchmark under a linear specification and reveals nonlinear variation in marginal returns under a richer structural class.

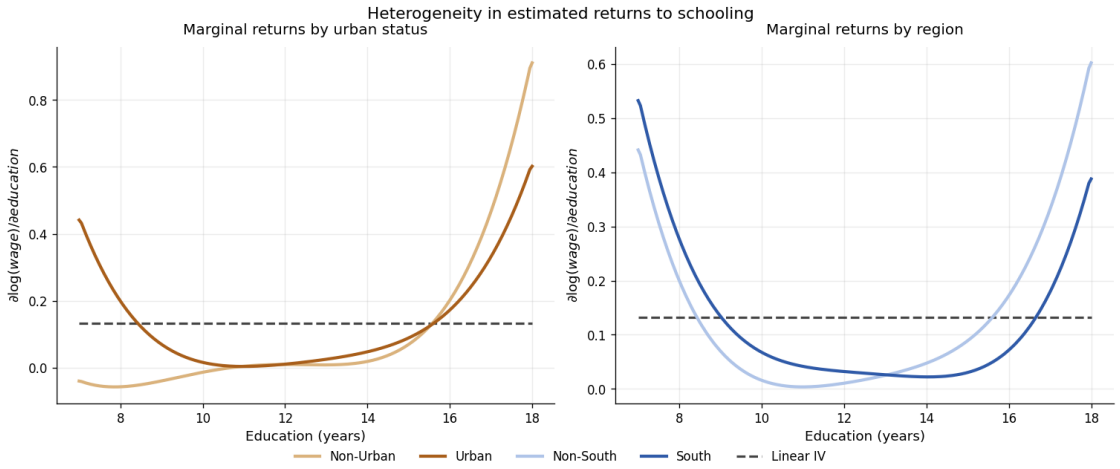


Figure 5.7: Model-implied heterogeneity in the nonlinear IV fit across observed environments. The figure reports subgroup-specific structural profiles together with the overall linear IV benchmark.

Future work should develop inference procedures that account for both the learned first stage and simulation error, extend the consistency theory to growing structural classes and trained diffusion models, and explore other conditional moment settings where generative models can be used as conditional integration devices.

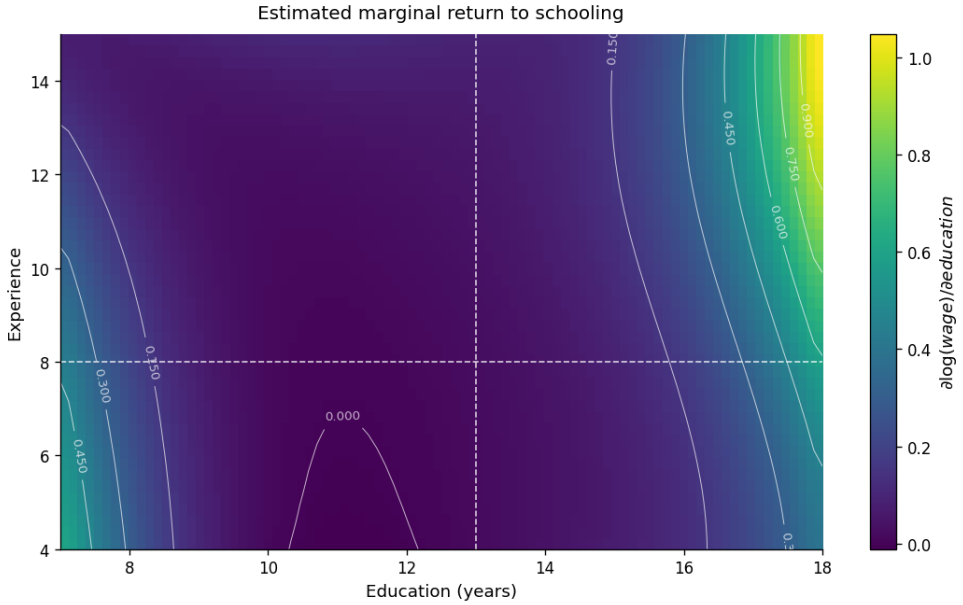


Figure 5.8: Estimated marginal return to schooling over schooling and labor-market experience. The heatmap plots the nonlinear IV estimate of $\partial g(x, w)/\partial x$ over schooling and experience, holding the remaining controls fixed at the baseline profile.

CHAPTER 6

GENERATIVE BOOTSTRAP AND LOCAL ENVIRONMENTS

The preceding chapters treated generative modeling as a way to represent an observed distribution. A sampler represents a distribution in operational form as it gives a way to draw from it. In the diffusion chapter, this representation was the object of analysis. In the generative IV chapter, it became a device for computation: by conditioning on a profile and simulating the remaining variables, the learned distribution made a conditional moment restriction usable.

This chapter records the broader implication. This chapter is intended to be an informal discussion and perhaps encouragement of potential future research directions. We believe that generative models should be viewed not only as machines for producing realistic synthetic data, but as possible instruments for simulation-based statistical methods. A learned sampler can stand in for an unknown distribution, or for a family of conditional distributions, and can be used to propagate statistical procedures through the fitted environment. Perhaps generative models fit a natural statistical role: they can supply surrogate distributions rich enough to support forms of simulation that are difficult to obtain from empirical or parametric approximations alone.

6.1 The Bootstrap

The natural point of contact is the bootstrap. The bootstrap is often introduced as resampling from the data. More generally, it is simulation from a surrogate distribution. The unknown data-generating distribution is replaced by an estimated distribution, and the statistic is recomputed under repeated draws from that estimate. In the nonparametric bootstrap, the surrogate is the empirical distribution

$$\widehat{P}_n = \frac{1}{n}\sum_{i=1}^{n} \delta_{Z_i}.$$

In the parametric bootstrap, it is a fitted distribution $P_{\widehat{\theta}}$. Other variants choose surrogates adapted to dependence, heteroskedasticity, clustering, or other features of the sampling environment. The common structure is the same: approximate the environment, simulate from it, and examine the statistic under the simulated distribution.

A generative bootstrap would replace the surrogate by a learned sampler. If $\widehat{P}_G$ denotes the distribution represented by a fitted generative model, then one may draw

$$Z_1^*, \ldots, Z_m^* \sim \widehat{P}_G$$

and recompute the statistic on the synthetic sample. The generator becomes a simulation mechanism through which an estimated distribution is made available.

This viewpoint serves as a potential starting point for a serious future research direction. The appeal of a generator is that it may provide a richer surrogate environment than the empirical distribution or a low-dimensional parametric family. It also remains functional–sampling enables Monte-Carlo calculation of a variety of statistical objects. However, a generative bootstrap must be justified relative to the functional it is meant to approximate. This requires a serious research treatment that involves theoretical rigor and empirical validation.

The classical bootstrap is powerful because it connects a computational recipe to an approximation statement. Generative bootstrap procedures require the same kind of discipline. For example, one must ask which discrepancy between $\widehat{P}_G$ and the true distribution matters, at what scale it must vanish, and how generator error interacts with sampling error and Monte Carlo error. Without such statements, simulation from a generator is useful as a diagnostic; with them, it may become an inferential method.

6.2 Conditional Simulation and Local Environments

An important case for this thesis is conditional simulation. Many statistical targets are defined not by a single marginal distribution, but by a family of conditional distributions,

$$P(\cdot \mid S = s).$$

A conditional generator estimates this family and allows the statistician to fix a profile $S = s$ while sampling the remaining variables. It gives operational access to a local part of the observed environment.

This was the role of the generator in Chapter 5. The structural function was characterized by

$$\mathbb{E}\{Y - g_0(X, W) \mid Z, W\} = 0.$$

At fixed values of (Z, W), the learned conditional distribution supplied draws of the variables entering the residual. Monte Carlo averages under that fitted distribution made the conditional moment criterion computable. The generator supplied the local simulation needed to evaluate the statistical restriction.

The same idea suggests a broader class of simulation-based procedures. Once $\widehat{P}_G(\cdot \mid S = s)$ has been learned, one can recompute moments, estimators, or diagnostics under repeated draws from the fitted local distribution. This can be used to examine where a criterion is stable, where support is weak, where heterogeneity appears, or where the fitted environment is doing most of the inferential work. In this sense, a conditional generator can make an observational environment inspectable.

The word local is essential. Conditional simulation does not intervene on the world. It does not create identification, repair lack of overlap, or convert association into causal structure. It gives access to a fitted conditional distribution. If the conditioning variables define a meaningful statistical comparison, simulation can make that comparison operational. If they do not, the generator does not supply the missing assumptions.

This is the right level at which to view generative models in many statistical problems: not as replacements for identification, but as tools that make identified or partially identified objects computationally accessible. They can enrich simulation-based inference, but only after the statistical target and the required approximation have been specified.

6.3 From Local Simulation to Stability

The role proposed here is deliberately forward-looking. Generative models offer a new source of surrogate distributions for simulation-based statistics. They may be useful for bootstrap procedures, conditional moment estimation, sensitivity analysis, diagnostics, and other settings where repeated sampling from a fitted environment clarifies the behavior of a statistic. Much of the necessary theory remains to be developed.

There is also a boundary to the idea. Simulation is local to the distribution that has been learned. It can reveal structure inside an observed environment, but it does not say which parts of that structure will persist when the environment changes. A relation may be stable under repeated simulation from the fitted distribution and still fail in a new population, time period, or deployment setting.

This boundary leads to the final part of the thesis. The next chapter studies prediction under concept shift. The question there is no longer how to represent an observed environment, nor how to simulate within it. The question is which predictive relations remain reliable across environments. The movement of the thesis is therefore from representation, to simulation, to stability.

CHAPTER 7

STABILIZING OBSERVED RELATIONS: CONFOUNDING, CONCEPT SHIFT, AND INVARIANT PREDICTION

7.1 Introduction

Practitioners often deploy predictive models in new environments where the covariate and response distributions have shifted. Consider two observational data sets collected from different environments: a source environment used for training and a target environment where the model is deployed. With observational data sets, a model maximizing prediction accuracy in the source environment may experience a substantial drop in accuracy in the target environment, often due to unobserved confounding factors lurking in the environments. Such confounding can fundamentally alter the data distribution across environments and, in turn, alter the underlying concept of which statistical model is best.

A predictive model that generalizes well to new data sets is the statistical ideal, yet theory and methodology weaken when the environment underpinning the data set shifts. Recently, a growing literature in causal inference has shown the utility of comparing data sets from distinct environments to identify causal relations [Peters et al., 2016, Heinze-Deml et al., 2017, Pfister et al., 2019, Arjovsky et al., 2019]. At the same time, a body of works in machine learning (ML) and artificial intelligence (AI) has been tackling prediction under distribution shifts, leading to a rich field called domain adaptation [Ben-David et al., 2006, 2010, Blitzer et al., 2006, 2007, Tzeng et al., 2014, Long et al., 2015, Ganin et al., 2016]. While both lines of research aim to address the problem of distribution shifts, they utilize two distinct notions of stability under different problem settings, which we discuss below.

Causal Stability Consider a covariate-response pair (X, Y) and its joint probability distribution $\mathcal{P}_{\mathcal{E}}(X, Y)$ that differs across environments $\mathcal{E}$. We consider two environments

$\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$ corresponding to source and target, respectively. Due to the presence of unobserved confounding factors, both the *covariate distribution* $\mathcal{P}_{\mathcal{E}}(X)$, the marginal distribution of X, and the *conditional concept* $\mathcal{P}_{\mathcal{E}}(Y|X)$, the conditional distribution of Y given X, could be shifted and altered by the change in the environment. The principle of independent mechanisms in causality [Peters et al., 2017] points to a plausible source of invariance: the mechanism producing the effect given the cause should remain unchanged across environments. This principle can be made practical. For instance, the invariant risk minimization work [Peters et al., 2016, Heinze-Deml et al., 2017, Pfister et al., 2019, Rothenhäusler et al., 2021] seeks to identify a causally "stable" representation, that is, a map $X \mapsto \phi(X)$ such that $\mathcal{P}(Y|\phi(X))$, the conditional concept given $\phi(X)$, stays invariant across environments (hence no subscript $\mathcal{E}$). This line of work requires labeled data across environments, specifically, it assumes access to samples from both $\mathcal{P}_{\mathcal{S}}(X, Y)$ and $\mathcal{P}_{\mathcal{T}}(X, Y)$.

Distributional Stability Unlike the setting above, domain adaptation considers the case where practitioners observe labeled covariate-response pairs from the source distribution $\mathcal{P}_{\mathcal{S}}(X, Y)$, but only *unlabeled* covariate samples from a target distribution $\mathcal{P}_{\mathcal{T}}(X)$. The seminal work of Ben-David et al. [2006] derived an upper bound on target risk, a key to domain adaptation. Given a representation $\phi : X \mapsto \phi(X)$, the upper bound involves a balancing act between (i) the source risk of the model under the representation ϕ, and (ii) a distance between probability distributions $\mathcal{P}_{\mathcal{S}}(\phi(X))$ and $\mathcal{P}_{\mathcal{T}}(\phi(X))$. Here, a notion of distributional "stability" arises. One prefers a representation $X \mapsto \phi(X)$ where the difference between $\mathcal{P}_{\mathcal{S}}(\phi(X))$ and $\mathcal{P}_{\mathcal{T}}(\phi(X))$ is small, holding source risk fixed. This notion of invariance is also natural: if the distribution shifts significantly, the model selected based on source data may suffer severely when extrapolating to target data.

This Chapter Studying concept shifts induced by unobserved confounding—a central topic in causal inference with observational data—has received comparatively less attention in domain adaptation. We propose a structural causal model to address the challenge of

domain adaptation in the presence of distribution shifts caused by unobserved confounding factors. This structural model unites notions of stability and the methods used to enforce them, developed independently in the ML/AI and statistics literatures.

Much like the role of instrumental variables in the presence of confounding, we find that leveraging an exogenous and invariant covariate representation can address both concept and covariate shifts lurking in the environments. Further, in the context of our structural model, notions of causal stability and distributional stability will align. With this understanding, we will motivate, from first principles, a stability-regularized risk minimization method that aims for invariance and stability when adapting to new observational domains.

7.1.1 A Structural Causal Model

Throughout the chapter, we postulate a Structural Causal Model (SCM) for unobserved confounding inspired by the idea of instrumental variables [Stock and Trebbi, 2003, Pearl, 2009]. The crucial difference is that, in our model, these exogenous variables—resembling instruments—are latent and unobserved.

Definition 7.1.1 (Model)**.** A pair of covariate and response, denoted as $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$, are generated from mutually independent exogenous random variables $E \in \mathbb{R}^r, Z \in \mathbb{R}^k$ with $r, k \leq d$ and $U \in \mathbb{R}, W \in \mathbb{R}^d$ via the following structural causal model,

$$Y = \langle \beta^\star, X \rangle + \langle \gamma, E \rangle + U \ , \tag{7.1.1}$$

$$X = \Theta Z + \Delta E + W \ , \tag{7.1.2}$$

where $\beta^\star \in \mathbb{R}^d$, $\gamma \in \mathbb{R}^r$, $\Theta \in \mathbb{R}^{d \times k}$, and $\Delta \in \mathbb{R}^{d \times r}$ are unknown parameters; see Figure 7.1.

The *unobserved confounding* variable E models the environmental influence, whose distribution shifts across settings. Concretely, in domain adaptation, the distribution of E shifts from the source distribution $\rho_{\mathcal{S}}$ to the target distribution $\rho_{\mathcal{T}}$, where $\rho_{\mathcal{S}}, \rho_{\mathcal{T}}$ are probability distributions defined on $\mathbb{R}^r$. The *latent invariant* Z models a source of randomness that is

not subject to environment shifts: a stable, exogenous structure shared across both source and target environments. It plays a similar role to an instrumental variable, with the notable difference that it is unobserved across environments. The *exogenous noise* U, W are mutually independent and mean-zero, and their distributions do not depend on the environment.

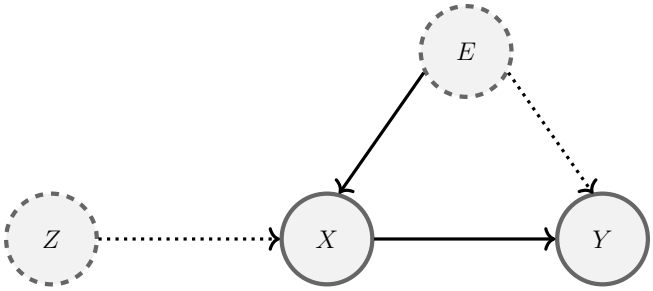


Figure 7.1: Diagram visualizing the model in Definition 7.1.1. The endogenous confounding variable E lurking in the environment and the exogenous invariant variable Z are both latent and unobserved.

Observational Data and the Adaptation Problem Consider the source and target environments $\mathcal{S}$ and $\mathcal{T}$. The joint distributions of (X, Y) under $E \sim \rho_{\mathcal{S}}$ and $E \sim \rho_{\mathcal{T}}$ are denoted as $\mathcal{P}_{\mathcal{S}}(X, Y)$ and $\mathcal{P}_{\mathcal{T}}(X, Y)$, with $\mathcal{P}_{\mathcal{S}}(X)$ and $\mathcal{P}_{\mathcal{T}}(X)$ denoting the covariate distributions. We use $\mathbb{E}_{\mathcal{E}}$ to denote the expectation w.r.t. the joint distribution under the environment $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$. For the expectation of random variables depending only on Z, U, W, we drop the subscript $\mathcal{E}$ as it is invariant across environments.

The practitioner observes labeled covariate-response pairs jointly drawn from the source distribution $\mathcal{P}_{\mathcal{S}}(X, Y)$, but only unlabeled covariate data from a target distribution $\mathcal{P}_{\mathcal{T}}(X)$. Given the observational data, the practitioner aims to estimate a linear model based on $\mathcal{P}_{\mathcal{S}}(X, Y)$ and $\mathcal{P}_{\mathcal{T}}(X)$ that performs well in the target environment.

Confounding and Concept Shift A few remarks follow for our model. First, suppose the dotted line connections in Figure 7.1 are removed. In that case, we arrive at the prototypical well-specified covariate shift setting: environment shifts only affect the covariate distribution, not the conditional distribution of Y given X. In this setting, without confounding, using (weighted) least squares to estimate a model from X to Y consistently is possible. Our model incorporating the dotted lines—capturing unobserved confounding—naturally extends the

covariate shift setting. Second, since the unobserved confounding variable E lurking in the environment influences both X and Y, the least squares estimate is biased and unreliable when deploying to new environments. As we shall see in Proposition 7.2.2, the distribution shift in E under this structural model induces a *concept shift*, meaning that the best linear model depends on the environment. Finally, introducing the invariant, exogenous Z may remind readers of instrumental variables. However, in our domain adaptation problem, Z is unobserved; we merely postulate its existence. We develop a methodology that leverages invariance across environments to learn a stable, lower-dimensional linear subspace without directly observing Z.

7.1.2 A Motivating Example

We peek at data from an influential economic study by Tabellini [2010]. Tabellini uses this data to explore the impact of historic social, economic and political development on the European economy in the late 90s. We will treat this data as observational, and consider the covariate-response pair:

$$X = \{\textit{Institutions}, \textit{Literacy}, \textit{Enrollment}, \textit{Culture}, \textit{Urbanization}\}, \; Y = \{\textit{Economic Output}\} \; .$$

Due to circumstances, suppose we only have access to data from rural regions (the source environment) and wish to construct a linear predictor for Y in historically urban areas (the target environment). How do we find an accurate predictor for *Economic Output* in the target environment? Considering our feature set, X, we would certainly expect distribution shifts in at least one coordinate—'Urbanization'. Furthermore, Tabellini establishes 'Institutions' as an instrumental variable that is independent of 'Urbanization', and so, we would also expect at least one feature to be stable amid this environment shift. We can capture this setting in our data model; 'Institutions' and 'Urbanization' are natural candidates as components in Z and E, respectively. However, we do not assume this structure a priori, and hope to leverage

the stability of certain features, like 'Institutions', in a data-driven way. We will revisit this example in Section 7.5, after introducing the stability-regularized risk minimization method.

Notations Let $\|v\|$ denote the standard Euclidean norm of a vector $v \in \mathbb{R}^d$. For any symmetric matrix $M \in \mathbb{R}^{d\times d}$, let $\lambda_{\max}(M), \lambda_{\min}(M)$ denote the largest, smallest eigenvalues of M, respectively; if M is positive semidefinite, let $M^{1/2}$ denote its matrix root and $\|v\|_M := \|M^{1/2}v\|$ for $v \in \mathbb{R}^d$. Also, $\|\cdot\|_{\mathrm{op}}, \|\cdot\|_{\mathrm{F}}$ are matrix norms (spectral, Frobenius, respectively). The Stiefel manifold is defined as $\mathrm{St}(d,\ell) := \{A \in \mathbb{R}^{d\times\ell} : \ A^\top A = I_\ell\}$. Finally, $R_{\mathcal{E}}$ denotes the squared risk depending on environment $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$, namely, $R_{\mathcal{E}}(\beta) = \mathbb{E}_{\mathcal{E}}[(Y - \langle X, \beta\rangle)^2]$, where the expectation is w.r.t. the joint distribution $\mathcal{P}_{\mathcal{E}}(X, Y)$.

7.1.3 Organization and Contribution

The chapter studies domain adaptation in the presence of unobserved confounding. Unobserved confounding is a likely issue for a practitioner attempting to extrapolate with observational data. As seen in Figure 7.1, we extend the standard covariate shift setting to incorporate confounding factors lurking in the environment. These latent variables simultaneously influence covariate X and response Y.

In Section 7.2, we explore the shortcomings of vanilla risk minimization when both covariate distribution and "concept" are expected to shift under the linear SCM. As shown in Proposition 7.2.4, an invariant subspace in this model will unify concept stability and distributional stability, two distinct ideas in the literature a priori. Proposition 7.2.6 further establishes the necessity and benefit of leveraging a lower-dimensional, invariant subspace for predictive power.

Guided by insights into the invariant subspace, we solve the domain adaptation task in Section 7.3. Given distributional access to the unlabeled target, $\mathcal{P}_{\mathcal{T}}(X)$, and labeled data from the source, $\mathcal{P}_{\mathcal{S}}(X, Y)$, we seek a methodology gauging towards the inaccessible target risk by learning a subspace that balances predictability and stability/invariance. Concretely, we seek an estimator composed of a linear subspace representation, $V \in \mathbb{R}^{d\times\ell}$, and a low-

dimensional ridge regressor, $\alpha \in \mathbb{R}^\ell$, jointly optimized according to the objective

$$V,\ \alpha := \arg\min_{V,\alpha} \mathbb{E}_{\mathcal{S}}[(Y - \langle X, V\alpha\rangle)^2] + \upsilon\|\alpha\|^2 + \tfrac{\eta}{2} \underbrace{\|V^\top(\mathbb{E}_{\mathcal{T}}[XX^\top] - \mathbb{E}_{\mathcal{S}}[XX^\top])V\|_{\mathrm{F}}^2}_{(*)} . \qquad (7.1.3)$$

We define the composed estimator $\beta^{\upsilon,\eta} = V\alpha$.

Proposition 7.3.3 derives (7.1.3) as an upper bound on the target risk. Therefore, the $(*)$ term is a natural notion of invariance for domain adaptation under the linear SCM. Importantly, (7.1.3) is an actionable proxy for optimization based on the information available, and in Section 7.3.2, a practical first-order manifold optimization method is devised to navigate its non-convex landscape. By optimizing over the choice of regularization parameters η, υ in (7.1.3), the target risk can effectively be optimized. This empirical fact is demonstrated using real-world data examples in Section 7.5.

Moving to Section 7.4, we provide a theoretical characterization of the landscape of the non-convex manifold optimization. When the regularization parameter η is sufficiently large, we show that almost all local optima align well with an exogenous, invariant linear subspace and are resilient to distribution shifts driven by endogenous confounding factors. Indeed, denoting $\mathrm{Endo} \subset \mathbb{R}^d$ as the endogenous subspace corresponding to Δ in (7.1.2), we find the first-order stationary point $V \in \mathrm{St}(d, \ell)$ of (7.1.3) satisfies:

$$\max_{\mathbf{v}\in V, \mathbf{w}\in \mathrm{Endo}} \big(\cos\angle(\mathbf{v}, \mathbf{w})\big)^6 \le O\left(\frac{1}{\upsilon\eta^2}\right) .$$

This alignment result leads to a stability bound between the target and source risks that directly addresses the domain adaptation problem. For any stationary point of (7.1.3),

$$R_{\mathcal{T}}(\beta^{\upsilon,\eta}) - R_{\mathcal{S}}(\beta^{\upsilon,\eta}) \le \inf_{\beta\perp\mathrm{Endo}} \underbrace{\{R_{\mathcal{T}}(\beta) - R_{\mathcal{S}}(\beta)\}}_{\text{oracle term}} + O\left(\frac{1}{\upsilon^{4/3}\eta^{2/3}}\right) .$$

Namely, a predictive model using the learned lower-dimensional subspace could incur a nearly ideal gap between target and source risk, quantified by the oracle term. The theoret-

ical results on invariance alignment and risk stability corroborate empirical observations in real-world data sets. Despite the difficult nature of non-convex manifold optimization, the practical first-order method often identifies good local optima for domain adaptation. Lastly, we also provide a finite sample error analysis to quantify the gap between the optimization and its plug-in estimation.

We provide a detailed literature review and discussion, deferred to Appendix C in the Supplementary Material, due to the space limit. All technical proofs are collected in Appendix C for the same reason.

7.2 Undesired Properties of Source Risk Minimization

In this section, we study the linear SCM in Definition 7.1.1, identify the undesired theoretical properties of vanilla source risk minimization, and further characterize the benefits of leveraging an invariant subspace to enforce stable learning and improve the target risk.

7.2.1 Concept Shift, Confounding, and Subspace Invariance

We first state the assumptions used in this section.

Assumption 7.2.1. In Definition 7.1.1, the exogenous variables Z, U, W are mean-zero, namely, $\mathbb{E}[Z] = 0, \mathbb{E}[U] = 0, \mathbb{E}[W] = 0$, and satisfy $\mathbb{E}[WW^\top] = \tau^2 I_d$ with $\tau > 0$.

For each environment $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$, classical risk minimization takes the following form:

$$\arg\min_{\beta \in \mathbb{R}^d} R_{\mathcal{E}}(\beta) = \arg\min_{\beta \in \mathbb{R}^d} \mathbb{E}_{\mathcal{E}}[(Y - \langle X, \beta \rangle)^2] \,. \tag{7.2.1}$$

In the SCM in Definition 7.1.1, when the endogeneity parameter $\gamma \neq 0$, the unobserved variable E lurking in the environment plays a confounding role in the relationship between X and Y. A natural question is: will there be a linear model simultaneously optimal for source and target risk minimization? The answer is no, thus implying that the best model

will be a moving target. We shall see that a shift in the distribution of the confounding variable $E \sim \rho_{\mathcal{S}}$ to $E \sim \rho_{\mathcal{T}}$ can lead to a stark difference in the best linear model across distributions. The following proposition formally defines the above phenomenon as a *concept shift*.

Proposition 7.2.2 (Concept Shift)**.** *Under Assumption 7.2.1, the risk minimization* (7.2.1) *admits a unique minimizer, denoted as* $\beta_{\mathcal{E}}$ *for* $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$*. Suppose* $[\Theta, \Delta] \in \mathrm{St}(d, k+r)$*. Then, if* $\mathbb{E}_{\mathcal{S}}[EE^\top] \neq \mathbb{E}_{\mathcal{T}}[EE^\top]$*, there exists* $\gamma \in \mathbb{R}^r$ *such that* $\beta_{\mathcal{S}} \neq \beta_{\mathcal{T}}$*, namely, the best linear predictor (concept) shifts across the two environments for some endogeneity parameter* γ*.*

Remark 7.2.3. It is immediate to show that under Assumption 7.2.1, the best linear model is well-defined and given as $\beta_{\mathcal{E}} = \beta^\star + (\mathbb{E}_{\mathcal{E}}[XX^\top])^{-1}\Delta\mathbb{E}_{\mathcal{E}}[EE^\top]\gamma$. The above result shows that the best linear model shifts as long as the environment changes. The curious reader may wonder whether the Bayes optimal model $\mathbb{E}_{\mathcal{E}}[Y|X]$ also changes. If we assume in addition that (E, Z, W) is drawn from a multivariate Gaussian and E is mean-zero, we can show $\mathbb{E}_{\mathcal{E}}[Y|X] = \langle \beta^\star + (\mathbb{E}_{\mathcal{E}}[XX^\top])^{-1}\Delta\mathbb{E}_{\mathcal{E}}[EE^\top]\gamma, X\rangle$, which matches the best linear model. Therefore, the Bayes optimal concept is also environment dependent. The simple derivation above also shows that the re-weighting method of Shimodaira [2000], based on the likelihood ratio of the covariate distributions, will not fix the concept shift issue. Due to endogeneity, a new methodology is needed.

In plain language, source risk minimization fails to recover a model that stays invariant across environments. The concept shift induced by the movement of the second moments of E reflects a bias that cannot be reconciled through traditional least squares. The root cause of the above concept shift is endogeneity. Inspired by the instrumental variable literature, we instead resort to exogenous linear subspaces of covariates invariant to environment shifts. The following proposition shows two desiderata—dimension reduction and invariance—can be achieved simultaneously.

Proposition 7.2.4 (Subspace Invariance)**.** *Under Assumption 7.2.1, let* $\beta_{\mathcal{E}}^{\Theta}$ *be the best linear predictor restricted to the linear subspace* $\Theta \in \mathrm{St}(d,k)$ *for each environment* $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$*:*

$$\beta_{\mathcal{E}}^{\Theta} := \Theta\alpha_{\mathcal{E}}^{\Theta}, \text{ where } \alpha_{\mathcal{E}}^{\Theta} = \underset{\alpha\in\mathbb{R}^k}{\arg\min}\, \mathbb{E}_{\mathcal{E}}[(Y - \langle X, \Theta\alpha\rangle)^2] \,.$$

Suppose $[\Theta, \Delta] \in \mathrm{St}(d, k+r)$*. Then, for any* $\gamma \in \mathbb{R}^r$*, we have* $\beta_{\mathcal{S}}^{\Theta} = \beta_{\mathcal{T}}^{\Theta}$*.*

Remark 7.2.5. The above result shows that deconfounding is possible with an invariant subspace. In our model, the effect of the confounding variable, E, is restricted to the linear subspace Δ. Therefore, the covariate distribution projected to a lower-dimensional linear subspace, Θ, remains stable despite the distribution shifts from $\mathcal{P}_{\mathcal{S}}(X,Y)$ to $\mathcal{P}_{\mathcal{T}}(X,Y)$. Proposition 7.2.4 should be read in contrast to Proposition 7.2.2; risk minimization constrained to this "invariant" linear subspace resists arbitrary confounding effects parametrized by γ. And so, restricting to the invariant, exogenous space Θ yields stability in learned concepts, resilient to environment shifts.

In view of the decomposition, $Y = \langle \Theta^\top\beta^\star, Z\rangle + \langle \Delta^\top\beta^\star + \gamma, E\rangle + \text{noise}$, the subspace model can be shown as $\beta_{\mathcal{E}}^{\Theta} \equiv \Theta\Theta^\top\beta^\star$. This decomposition separates variation in Y into invariant, exogenous component $\langle \Theta^\top\beta^\star, Z\rangle$ and environment dependent, endogenous component $\langle \Delta^\top\beta^\star + \gamma, E\rangle$. In particular, when $(I - \Theta\Theta^\top)\beta^\star = 0$, learning in the exogenous space consistently recovers the true signal; $\beta_{\mathcal{E}}^{\Theta} = \beta^\star$.

Proposition 7.2.4 only demonstrates the existence of an invariant linear subspace Θ. However, since the exogenous variable Z is unobserved, a data-driven approach to learning such a lower-dimensional linear subspace is still largely unclear. Motivated by invariance and stability, in Section 7.3, we will develop a methodology that inherits a manifold optimization problem to learn a lower-dimensional linear subspace. Later in Section 7.4, we will derive a theoretical characterization of the subspace alignment between the invariant Θ and a first-order stationary point of the manifold optimization. A target risk bound exploiting this alignment will also be established therein.

7.2.2 Target Risk Improvement via Invariance

For domain adaptation, the lingering question is whether leveraging the invariant subspace Θ can improve the target risk compared to $\beta_{\mathcal{S}}$, the vanilla source risk minimizer. This section provides a sufficient and necessary condition for the above question. The following proposition delineates when one can simultaneously achieve three desiderata: target risk improvement, invariance, and dimension reduction.

Proposition 7.2.6 (Target Risk Improvement)**.** *Denote $\Lambda_{\mathcal{E}} := \mathbb{E}_{\mathcal{E}}[EE^\top]$ for $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$. Under Assumption 7.2.1, and the assumption that $[\Theta, \Delta] \in \mathrm{St}(d, k+r)$ with $k+r=d$. Then, the linear subspace predictor indexed by Θ based on source risk, that is, $\beta_{\mathcal{S}}^{\Theta}$, has a smaller target risk than the usual source risk minimizer $\beta_{\mathcal{S}}$, namely, $R_{\mathcal{T}}(\beta_{\mathcal{S}}^{\Theta}) < R_{\mathcal{T}}(\beta_{\mathcal{S}})$ if and only if*

$$\|\Delta^\top \beta^\star + (\tau^2 I_r + \Lambda_{\mathcal{T}})^{-1}\Lambda_{\mathcal{T}}\gamma\|^2_{\tau^2 I_r + \Lambda_{\mathcal{T}}} < \|(\tau^2 I_r + \Lambda_{\mathcal{S}})^{-1}\Lambda_{\mathcal{S}}\gamma - (\tau^2 I_r + \Lambda_{\mathcal{T}})^{-1}\Lambda_{\mathcal{T}}\gamma\|^2_{\tau^2 I_r + \Lambda_{\mathcal{T}}} .$$

The right-hand side can be interpreted as the amount of concept shift, and the left-hand side can be interpreted as the sub-optimality of target risk due to the (invariant) subspace model. A special case helps to unpack the result. Consider $(I - \Theta\Theta^\top)\beta^\star = 0$. Then, the above expression reads

$$\|\Lambda_{\mathcal{T}}\gamma\|_{(\tau^2 I_r + \Lambda_{\mathcal{T}})^{-1}} < \|\tau^2 \underbrace{(\Lambda_{\mathcal{T}} - \Lambda_{\mathcal{S}})}_{\text{environment shift}} (\tau^2 I_r + \Lambda_{\mathcal{S}})^{-1}\gamma\|_{(\tau^2 I_r + \Lambda_{\mathcal{T}})^{-1}} .$$

Conceptually, when the environment shift—parametrized by the second moment of the confounding variable E—is large enough, the subspace model strictly improves upon the vanilla source risk minimizer. When such a condition holds, target risk improvement, invariance, and dimension reduction can be achieved simultaneously.

Before concluding this section, we use the following simple example to elucidate the condition in Proposition 7.2.6.

Example 7.2.7. Consider the simple setting where $\Lambda_{\mathcal{S}} = \sigma_s^2 \cdot I_r$, $\Lambda_{\mathcal{T}} = \sigma_t^2 \cdot I_r$, and $\Delta^\top \beta^\star = c \cdot \gamma$ with a scalar $c \in \mathbb{R}$. Proposition 7.2.6 boils down to: $R_{\mathcal{T}}(\beta_{\mathcal{S}}^{\Theta}) < R_{\mathcal{T}}(\beta_{\mathcal{S}})$ if and only if

$$\left(c + \frac{\sigma_t^2}{\sigma_t^2 + \tau^2} \right)^2 < \left(\frac{\sigma_s^2}{\sigma_s^2 + \tau^2} - \frac{\sigma_t^2}{\sigma_t^2 + \tau^2} \right)^2 .$$

Fixing other parameters σ_t, σ_s, τ, this reduces to inequalities for the scalar parameter c; (i) *Target Rich Regime* with $\sigma_t > \sigma_s$: $\frac{-\sigma_s^2}{\sigma_s^2+\tau^2} + \frac{-2\tau^2(\sigma_t^2-\sigma_s^2)}{(\sigma_t^2+\tau^2)(\sigma_s^2+\tau^2)} < c < \frac{-\sigma_s^2}{\sigma_s^2+\tau^2}$; (ii) *Source Rich Regime* with $\sigma_s > \sigma_t$: $\frac{-\sigma_s^2}{\sigma_s^2+\tau^2} < c < \frac{-\sigma_s^2}{\sigma_s^2+\tau^2} + \frac{-2\tau^2(\sigma_t^2-\sigma_s^2)}{(\sigma_t^2+\tau^2)(\sigma_s^2+\tau^2)}$. In summary, the magnitude of confounding and the amount of environment shift altogether determine when invariance renders risk improvement.

So far, we have shown that leveraging subspace invariance allows one to surpass the limitations of source risk minimization in situations where the concept and the covariate shift. In the next section, we will propose a new domain adaptation method that learns a linear subspace resilient to shifting environments, making the insights obtained in this section actionable. The method trades off predictability and stability, using a lower-dimensional representation as a vehicle.

7.3 Methodology: Domain Adaptation via Manifold Optimization

The previous section briefly discusses the fundamental tradeoff between invariance and predictive performance. In general, an invariant subspace can improve upon the source risk minimizer but may not be optimal for the target risk in domain adaptation. This section contributes to delineating this tradeoff in the SCM defined in Definition 7.1.1. Concretely, we design a representation learning method for domain adaptation, parametrized by subspace projections, and navigate the tradeoff between representation invariance and risk minimization. Though the proposed manifold optimization is non-convex, we demonstrate that a standard first-order method works both empirically in Section 7.5 with real data sets, and later provably in Section 7.4, where we develop theory matching the empirics.

We motivate our main methodology in two ways: (i) as a surrogate for an upper bound on the target risk directly, and (ii) as a tradeoff to balance robustness/invariance and predictive performance.

7.3.1 A Surrogate for the Target Risk

In typical domain adaptation, labels from the target environment $Y \sim \mathcal{P}_{\mathcal{T}}(Y)$ are unavailable, and thus $R_{\mathcal{T}}(\beta)$ cannot be directly minimized. To circumvent this issue, we design a simple surrogate objective that does not require labels from the target environment, requiring only information about $\mathcal{P}_{\mathcal{S}}(X, Y)$ and $\mathcal{P}_{\mathcal{T}}(X)$. In the sequel we will denote $\Sigma_{\mathcal{E}} := \mathbb{E}_{\mathcal{E}}[XX^\top]$ for $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$. We start with a simple algebraic relation between the source and the target.

Proposition 7.3.1. *Under Assumption 7.2.1 and $\Delta^\top \Delta = I_r$, we have*

$$R_{\mathcal{T}}(\beta) = R_{\mathcal{S}}(\beta) + \left\langle \beta - (\beta^\star + \Delta\gamma), (\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})(\beta - (\beta^\star + \Delta\gamma)) \right\rangle \quad \forall \beta \in \mathbb{R}^d \,. \tag{7.3.1}$$

The geometry of the target risk landscape involves a balance of two terms with clear distance interpretations. The first term $R_{\mathcal{S}}(\beta)$ is the source risk. The second term encodes a distance between β and the endogenous component, $\beta^\star + \Delta\gamma$, under a geometry determined by the covariance shift matrix $\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}}$. The second term is not actionable based on data because $\beta^\star + \Delta\gamma$ is unknown. In the following proposition, we define an actionable objective, which serves as a surrogate upper bound for the target risk, for all β and $\beta^\star + \Delta\gamma$.

To make the presentation simple, we impose the following additional assumption.

Assumption 7.3.2 (Target Richer than Source)**.** Let $\Lambda_{\mathcal{E}} := \mathbb{E}_{\mathcal{E}}[EE^\top]$ for $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$. Assume that $D := \Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}} \succ 0$.

Note that the assumption is for simplicity of presentation: we can state more general results by separating the positive and negative parts of $\Lambda_{\mathcal{T}} - \Lambda_{\mathcal{S}}$. More importantly, we use Assumption 7.3.2 to focus on the interesting regime of distribution shift where the target

distribution strictly dominates the source. This is when out-of-domain extrapolation can happen. Informally speaking, the Loewner order, $\Lambda_{\mathcal{T}} \succ \Lambda_{\mathcal{S}}$, indicates more variability present in the target domain.

With Assumption 7.3.2 in hand, an upper bound on target risk can be derived, which suggests a practical proxy for $R_{\mathcal{T}}(\beta)$. To elucidate the low-dimensional subspace dependence, we explicitly construct the estimator β to be a composite map of a lower-dimensional representation, parametrized by $V \in \mathbb{R}^{d\times\ell}$, and a linear model restricted to the representation, parametrized by $\alpha \in \mathbb{R}^{\ell}$.

Proposition 7.3.3 (Surrogate for Target)**.** *If Assumptions 7.2.1 and 7.3.2 hold with* $\Delta^\top\Delta = I_r$*, then for any* $(V,\alpha) \in \mathbb{R}^{d\times\ell}\times\mathbb{R}^{\ell}$ *and* $\xi,\zeta > 0$*, we have* $R_{\mathcal{S}}(V\alpha) \le R_{\mathcal{T}}(V\alpha)$ *and*

$$R_{\mathcal{T}}(V\alpha) \le R_{\mathcal{S}}(V\alpha) + \frac{(1+\xi)\zeta}{2}\|\alpha\|^4 + \frac{(1+\xi)}{2\zeta}\|V^\top DV\|_{\mathrm{F}}^2 + (1+\tfrac{1}{\xi})\langle\beta^\star+\Delta\gamma, D(\beta^\star+\Delta\gamma)\rangle\ . \quad (7.3.2)$$

Proposition 7.3.3 highlights that the gap between target and source risks of a composite estimator $\beta = V\alpha$ can be controlled by two complexities: $\|\alpha\|$ and $\|V^\top(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})V\|_{\mathrm{F}}$. The term $\|V^\top(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})V\|_{\mathrm{F}}$ quantifies the alignment of the subspace V with the directions of covariance shift; in essence, a term describing stability for extrapolation. The term may also be interpreted as a quadratic maximum mean discrepancy (MMD) between covariate distributions in the linear subspace V. Indeed, under the map $\psi_V(X) = V^\top XX^\top V$,

$$\left\|V^\top(\mathbb{E}_{\mathcal{T}}[XX^\top] - \mathbb{E}_{\mathcal{S}}[XX^\top])V\right\|_{\mathrm{F}} = \sup_{M:\|M\|_{\mathrm{F}}\le 1} \mathbb{E}_{\mathcal{T}}[\langle M, \psi_V(X)\rangle] - \mathbb{E}_{\mathcal{S}}[\langle M, \psi_V(X)\rangle]\ . \quad (7.3.3)$$

Finally, note that the right-hand side of (7.3.2) is a regularized source risk minimization, regularized by ridge and stability penalties. More importantly, it is an actionable surrogate objective for domain adaptation, solely depending on $\mathcal{P}_{\mathcal{S}}(X,Y)$ and $\mathcal{P}_{\mathcal{T}}(X)$.

7.3.2 An Optimization Procedure on the Stiefel Manifold

Motivated by the upper bound on the target risk shown in Proposition 7.3.3, we define the following penalized objective to search for a linear subspace V, and a model restricted to that subspace parametrized by α,

$$F_{\upsilon,\eta}(V,\alpha) := \tfrac{1}{2}\left\{R_{\mathcal{S}}(V\alpha) + \upsilon\|\alpha\|^2 + \tfrac{\eta}{2}\|V^\top(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})V\|_{\mathrm{F}}^2\right\} . \tag{7.3.4}$$

By optimizing over a linear subspace $V \in \mathbb{R}^{d\times\ell}$ that balances predictive power and stability, the above objective provides an explicit tradeoff. Note that the two-level optimization is convex in α for fixed V, where the inner optimum $\min_{\alpha\in\mathbb{R}^\ell} F_{\upsilon,\eta}(V,\alpha)$ is attained at

$$\alpha_V := (V^\top \mathbb{E}_{\mathcal{S}}[XX^\top]V + \upsilon I_\ell)^{-1} V^\top \mathbb{E}_{\mathcal{S}}[XY] . \tag{7.3.5}$$

We search for a representation confined to the Stiefel manifold $V \in \mathrm{St}(d,\ell)$. Putting things together, we arrive at the following optimization on the Stiefel manifold,

$$\min_{V\in\mathrm{St}(d,\ell)} \Phi_{\upsilon,\eta}(V) := \min_{V\in\mathrm{St}(d,\ell)} F_{\upsilon,\eta}(V,\alpha_V) \tag{7.3.6}$$

where α_V is defined in (7.3.5). The problem (7.3.6) is non-convex in V; however, we will derive provable results for any first-order stationary point of the landscape. Informally speaking, as we shall see in Section 7.4, under certain conditions, almost all local optima demonstrate good alignment with the invariant subspace.

Let us first introduce a skeletal implementation of a first-order manifold optimization method to find local optima of $\Phi_{\upsilon,\eta}(V)$. Then, we move to fill in the essential background on the geometry of the Stiefel manifold to rationalize the algorithm.

Geometry on the Stiefel Manifold A few essential geometric items must be defined to understand the first-order method optimizing (7.3.6) detailed in Algorithm 7.3.1. We treat

Algorithm 7.3.1 Optimization on Stiefel Manifold

Require: Initial point: $V_0 \in \mathrm{St}(d, \ell)$.
  for $k = 0, 1, 2, \ldots$ **do**
    Gradient: $G_k \leftarrow (I - V_k V_k^\top)((\Sigma_{\mathcal{S}} V_k \alpha_{V_k} - \mathbb{E}_{\mathcal{S}}[XY])\alpha_{V_k}^\top + \eta D V_k V_k^\top D V_k)$ as per (7.3.7).
    Retraction: $V_{k+1}(t) := (V_k + t \cdot G_k)(I_\ell + t^2 \cdot G_k^\top G_k)^{-1/2}$ for $t \in \mathbb{R}_+$ as per (7.3.8).
    Line-Search: choose step-size $t = t_k$ via line-search, and set $V_{k+1} \leftarrow V_{k+1}(t_k)$.
  end for

$\mathrm{St}(d, \ell)$ as an embedded sub-manifold in the vector space $\mathbb{R}^{d \times \ell}$ and introduce appropriate notions of tangent space, projection, and gradient. We provide a cursory overview of the Stiefel geometry for a self-contained treatment. A familiar reader may skip this part.

First, the tangent space at a point on a sub-manifold is intuitively the plane tangent to the sub-manifold at that point. The normal space is the corresponding orthogonal complement. The tangent space and the normal space to $\mathrm{St}(d, \ell)$ at V are given as $T_V \mathrm{St}(d, \ell) = \left\{ Z \in \mathbb{R}^{d \times \ell} : V^\top Z + Z^\top V = 0 \right\}$ and $(T_V \mathrm{St}(d, \ell))^\perp = \left\{ VS : S = S^\top \in \mathbb{R}^{\ell \times \ell} \right\}$, respectively. We view the Stiefel manifold as a sub-manifold of $\mathbb{R}^{d \times \ell}$, with its tangent spaces inheriting the Euclidean metric (Frobenius norm) $\|\cdot\|_{\mathrm{F}}$ and inner product $\langle A, B \rangle = \mathrm{Tr}(A^\top B)$. Then, projections of $\xi \in \mathbb{R}^{d \times \ell}$ on to the tangent and the normal spaces at V are given by $P_V(\xi) = \left(I_d - VV^\top\right)\xi + V \mathrm{skew}\left(V^\top \xi\right)$ and $P_V^\perp(\xi) = V \mathrm{sym}\left(V^\top \xi\right)$, respectively, where $\mathrm{skew}(A) = (A - A^\top)/2$ and $\mathrm{sym}(A) = (A + A^\top)/2$.

Finally, if F is a smooth function defined on $\mathbb{R}^{d \times \ell}$, and $\bar{F}$ is its restriction to the Riemannian sub-manifold $\mathrm{St}(d, \ell)$, the gradient of $\bar{F}$ is equal to the projection of the gradient of F onto $T_V \mathrm{St}(d, \ell)$, namely, $\mathrm{grad}\ \bar{F}(V) = P_V(\mathrm{grad}\ F(V))$. A rigorous treatment of the above can be found in Absil et al. [2008, Chapters 3 and 4].

Gradient Formula and Retraction With the geometry understood, the Riemannian gradient of the objective (7.3.4) can be readily calculated.

Proposition 7.3.4 (Riemannian Gradient and Retraction)**.** *Consider* $\bar{\Phi}_{\upsilon,\eta} : \mathrm{St}(d, \ell) \to \mathbb{R}$*, the restriction of* $\Phi_{\upsilon,\eta}$ *to the Riemannian sub-manifold* $\mathrm{St}(d, \ell)$*. The Riemannian gradient*

of the objective (7.3.6) *is*

$$\text{grad } \bar{\Phi}_{\upsilon,\eta}(V) = (I_d - VV^\top)\left((\Sigma_{\mathcal{S}} V \alpha_V - \mathbb{E}_{\mathcal{S}}[XY])\alpha_V^\top + \eta DVV^\top DV\right) \in T_V \mathrm{St}(d,\ell)\ . \tag{7.3.7}$$

Moreover, the gradient formula for the objective (7.3.4) *is the same for either the Stiefel or the Grassmann manifold.*

For any $\xi \in T_V \mathrm{St}(d,\ell)$, the retraction map via matrix polar factorization is defined as

$$\mathrm{R}_V(\xi) = (V+\xi)(I_\ell + \xi^\top \xi)^{-1/2}\ . \tag{7.3.8}$$

Proposition 7.3.4 allows us to define first-order descent methods to optimize (7.3.6). For example, a simple line-search method on the Stiefel manifold updates the iterate as $V_{k+1} = \mathrm{R}_{V_k}(t_k \cdot G_k)$ for $G_k = \text{grad } \bar{\Phi}_{\upsilon,\eta}(V_k)$, where R_{V_k} is the retraction map from (7.3.8) and t_k is a step-size. Now, all ingredients of Algorithm 7.3.1 readily unfold.

7.3.3 Stability, Predictability and Connections to Modern ML

It was suggested in Section 7.3.2 that the penalized objective (7.3.6) is an instantiation of a robustness and accuracy tradeoff. Observe that under the linear SCM in Definition 7.1.1, two notions of stability/invariance coincide: optimizing over a *stable representation* amidst covariate shifts will align with searching for the *causal, invariant relationship*, stable across environments. We make the stability vs. predictability tradeoff explicit now and, by doing so, relate the method to a class of procedures recently popularized in ML. Our procedure seeks a two-part solution: a linear projection V, enforcing a similarity between $\mathcal{P}_{\mathcal{S}}$ and $\mathcal{P}_{\mathcal{T}}$, and a linear estimator, α, built atop. Selecting penalization then translates to a balancing of stability and predictability. This is transparent once we decouple $F_{\upsilon,\eta}$ into loss and

regularization,

$$F_{v,\eta}(V,\alpha) := \overbrace{\tfrac{1}{2}\mathbb{E}_{\mathcal{S}}[(Y-\langle X, V\alpha\rangle)^2]}^{\textit{Predictivity}} \quad + \quad \overbrace{\tfrac{v}{2}\|\alpha\|^2 + \tfrac{\eta}{4}\left\|V^\top(\Sigma_{\mathcal{T}}-\Sigma_{\mathcal{S}})V\right\|_{\mathrm{F}}^2}^{\textit{Stability}}.$$

The source *Predictivity* term corresponds to prediction under squared loss for the regression $Y \sim V^\top X$ confined to a subspace V; it extracts a linear relationship α between the label Y and a predictive subspace of the data $V^\top X$.

The *Stability* term quantifies two things. First, a notion of distributional stability determined by the penalty $\eta\left\|V^\top(\Sigma_{\mathcal{T}}-\Sigma_{\mathcal{S}})V\right\|_{\mathrm{F}}^2$ for the representation subspace V. The penalty can be interpreted as a distribution distance; see (7.3.3). Second, a standard ridge penalization quantifies the stability of the linear relationship α given the subspace V.

Conceptually, as $\eta \to \infty$, V captures the linear subspace over which the covariate second moments stay invariant. Moreover, such an invariant linear subspace corresponds to the contribution from the exogenous Z that identifies the invariant causal relationship. Indeed, $\Theta^\top\left(\mathbb{E}_{\mathcal{T}}[XX^\top]-\mathbb{E}_{\mathcal{S}}[XX^\top]\right)\Theta = 0$, therefore provided $l > k$, $\Theta \in \mathbb{R}^{d\times k}$ is an element of the non-empty solution set. In this sense, the case when $\eta \to \infty$ coincides with the invariant estimator explored in Proposition 7.2.6. More generally, as we will see in Section 7.5, a choice of η in a certain interval may balance stability and source predictability to harness improvements in target risk, the primary object in domain adaptation.

It is easy to see how the framework can be generalized to deal with different notions of distributional discrepancy, loss function, and model class. Specifically, a family of methods is based on finding a stable and predictive representation $\phi\colon \mathcal{X} \to \mathcal{Z}$, where $\mathcal{Z}$ is a suitable feature space. Let $\phi_{\#}\mu$ and $\phi_{\#}\nu$ denote the source and target covariate distributions pushed forward to the mapped space $\mathcal{Z}$. The goal is to find a representation $\phi\colon \mathcal{X} \to \mathcal{Z}$ and a composite predictor $g\colon \mathcal{Z} \to \mathcal{Y}$ simultaneously by solving $\min_{g,\phi} R_{\mathcal{S}}(g\circ\phi) + \eta \cdot d(\phi_{\#}\mu, \phi_{\#}\nu)$, where $d(\cdot,\cdot)$ is a suitable discrepancy between probability distributions on the mapped space $\mathcal{Z}$. Several choices for d have been proposed in the ML literature: Tzeng et al. [2014] and

Long et al. [2015] use the kernel MMD with a suitable kernel on $\mathcal{Z}$, Ganin et al. [2016] uses a suitable hypothesis-based discrepancy introduced in Ben-David et al. [2006, 2010], and Shen et al. [2018] uses the Wasserstein-1 distance, to name a few.

7.4 Theory: Invariance Alignment and Risk Stability

We now study theoretical properties of the methodology proposed in Section 7.3: alignment with the invariant subspace and risk across environments. The first result in Theorem 7.4.1 characterizes the landscape of the non-convex manifold optimization. We show almost all local optima exhibit favorable alignment with the invariant linear subspace as long as the regularization parameter is sufficiently large. The second result speaks to domain adaptation. Building on the landscape result, we derive a stability bound between the source and target risk in Theorem 7.4.3, which provides theoretical underpinnings for the empirical phenomena observed in Section 7.5. On top of these, Theorem 7.4.5 analyzes a finite-sample estimation error by quantifying the gap between the optimization procedure (7.3.6) and its plug-in estimation.

7.4.1 Alignment with the Invariant Subspace

In Propositions 7.2.4 and 7.2.6, the invariant subspace Θ is shown to be advantageous. The following theorem proves that the data-driven manifold optimization method given in Section 7.3 can identify linear subspaces that are approximately orthogonal to the contribution of the endogenous, confounding variable E. We employ the canonical correlation (or angle) between linear subspaces to quantify the notion of approximate orthogonality. By dodging the endogenous subspace where E influences X, the learned subspace V approximately aligns with the invariant subspace—crucial for concept invariance and distributional stability.

Theorem 7.4.1 (Alignment with the Invariant Subspace)**.** *Consider the model in Definition 7.1.1 that satisfies Assumptions 7.2.1 and 7.3.2, and* $\Delta^\top \Delta = I_r$*. Recall that* $D :=$

$\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}}$ *denotes the matrix of covariance shift. Suppose* $V \in \mathrm{St}(d, \ell)$ *satisfies the following conditions:*

1. V *is a first-order stationary point of the optimization* (7.3.6) *with* $\mathrm{grad}\ \Phi_{\upsilon,\eta}(V) = 0$,
2. V *is in the admissible set* $\mathcal{S}_\Delta(\delta)$, *where* $\mathcal{S}_\Delta := \{V \in \mathrm{St}(d, \ell)\ :\ \|V^\top \Delta\|_{\mathrm{op}}^2 \leq 1 - \delta\}$.

Then, we have

$$\|V^\top \Delta\|_{\mathrm{op}}^6 \leq \frac{\delta^{-1}\lambda_{\max}(\Sigma_{\mathcal{S}})\|\Sigma_{\mathcal{S}}^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^4}{\lambda_{\min}(\Delta^\top D \Delta)^4} \frac{1}{4\upsilon\eta^2}\ .$$

Remark 7.4.2. Theorem 7.4.1 characterizes the optimization landscape on the Stiefel manifold and should be interpreted when the regularization parameter η is large. It has two noteworthy features. First, we establish a structural result for *all local optima* for the non-convex manifold optimization, which can be found efficiently in practice using Algorithm 7.3.1; we do not require the solution V to be the global optima. Second, the result operates even when the invariant subspace overlaps with the endogenous space, namely $\Theta^\top \Delta \neq 0$. In such a case, the learned subspace V will still be approximately orthogonal to Δ, in the sense that $\max_{\mathbf{v}\in V, \mathbf{w}\in\Delta} \cos\angle(\mathbf{v}, \mathbf{w}) \leq \text{const.} \times \frac{1}{(\upsilon\eta^2)^{1/6}}$. Conceptually, the optimization procedure will identify a strict subspace inside the invariant subspace to approximately dodge the endogenous subspace Δ, so long as the stability regularization parameter η is large and ridge regularization parameter υ is not too small.

7.4.2 Stability and Target Risk Bound

Given the alignment with the invariant subspace, what remains to be shown is its implication for the domain adaptation problem, a task for this section. In that regard, the following theorem derives a stability bound between the target and the source risks. Recall the estimator $\beta^{\upsilon,\eta} = V\alpha_V$ with α_V as in (7.3.5). The following theorem shows the stability of any stationary point V of the non-convex Stiefel manifold optimization (7.3.6).

Theorem 7.4.3 (Stability between Source and Target)**.** *Consider the setting as in Theorem 7.4.1. Let* $V \in \mathrm{St}(d, \ell)$ *be any first-order stationary point of the optimization procedure*

(7.3.6) *with regularization parameters* υ, η, *and* $\beta^{\upsilon,\eta}$ *be the corresponding linear subspace estimator restricted to* V. *For any* $\epsilon > 0$, *define*

$$\mathsf{S}_{\epsilon,\delta} := (1+\epsilon^{-1})\delta^{-1/3}\lambda_{\max}(\Sigma_{\mathcal{S}})^{1/3}\frac{\lambda_{\max}(\Delta^\top D\Delta)}{\lambda_{\min}(\Delta^\top D\Delta)^{4/3}}\|\Sigma_{\mathcal{S}}^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^{10/3} ,$$

the following generalization bound holds

$$R_{\mathcal{T}}(\beta^{\upsilon,\eta}) - R_{\mathcal{S}}(\beta^{\upsilon,\eta}) \leq (1+\epsilon)\cdot\langle\beta^\star + \Delta\gamma, D(\beta^\star+\Delta\gamma)\rangle + \mathsf{S}_{\epsilon,\delta}\cdot\frac{1}{(4\upsilon)^{4/3}\eta^{2/3}} .$$

Remark 7.4.4. Note that the first term is necessary. Even for the oracle V that is perfectly invariant to the environment shift, the oracle gap between the target and source risks equals $\langle\beta^\star+\Delta\gamma, D(\beta^\star+\Delta\gamma)\rangle$, in view of Proposition 7.3.1. The above result proves an approximate oracle to the gap between target and source risk, quantified by ϵ. As $\eta\to\infty$ (holding all else fixed), the stability bound $S_{\epsilon,\delta}\cdot\frac{1}{(4\upsilon)^{4/3}\eta^{2/3}}$ decreases with η, which confirms the numerical findings in Section 7.5. Conceptually, as η increases, the source risk $R_{\mathcal{S}}(\cdot)$ will increase as one trades off prediction power for stability; on the plus side, instability $R_{\mathcal{T}}(\cdot) - R_{\mathcal{S}}(\cdot)$ is reduced due to invariance.

7.4.3 Finite-Sample Error Analysis

In practice, we have access to labeled and unlabeled samples from the source and target domains, respectively. Accordingly, we approximate the optimization (7.3.6) by replacing $\Phi_{\upsilon,\eta}$ with the plug-in analog $\widehat{\Phi}_{\upsilon,\eta}$; see (7.4.1) below. A natural statistical question is whether the plug-in estimator $\widehat{\Phi}_{\upsilon,\eta}$ is a good approximation to the true objective function $\Phi_{\upsilon,\eta}$. To answer this, we bound the uniform deviation of the plug-in estimator $\widehat{\Phi}_{\upsilon,\eta}$ from the true objective function $\Phi_{\upsilon,\eta}$ over the Stiefel manifold $\mathrm{St}(d,\ell)$. This allows us to quantify the gap between the objective value $\Phi_{\upsilon,\eta}$ evaluated at the empirical minimizer, namely, a minimizer of $\widehat{\Phi}_{\upsilon,\eta}$, and the minimum of the population optimization (7.3.6).

Theorem 7.4.5. *Let $\{(x_i, y_i)\}_{i=1}^n$ and $\{\tilde{x}_i\}_{i=1}^n$ be independent samples from $\mathcal{P}_\mathcal{S}(X, Y)$ and $\mathcal{P}_\mathcal{T}(X)$, respectively. Suppose that the supports of $\mathcal{P}_\mathcal{S}(X, Y)$ and $\mathcal{P}_\mathcal{T}(X, Y)$ are contained in $\{(x, y) \in \mathbb{R}^d \times \mathbb{R} : \|x\| \leq M, |y| \leq M\}$ for some $M > 0$. Let $\widehat{\Phi}_{\upsilon,\eta}$ be the plug-in estimator of $\Phi_{\upsilon,\eta}$ based on the samples: with the sample covariance matrices $\widehat{\Sigma}_\mathcal{T}, \widehat{\Sigma}_\mathcal{S}$, define*

$$\widehat{\Phi}_{\upsilon,\eta}(V) := \min_{\alpha \in \mathbb{R}^\ell} \tfrac{1}{2}\{\tfrac{1}{n}\sum_{i=1}^n (y_i - \langle V\alpha, x_i\rangle)^2 + \upsilon\|\alpha\|^2 + \tfrac{\eta}{2}\|V^\top(\widehat{\Sigma}_\mathcal{T} - \widehat{\Sigma}_\mathcal{S})V\|_{\mathrm{F}}^2\} \,. \tag{7.4.1}$$

Let $\widehat{V} \in \mathrm{St}(d, \ell)$ be a minimizer of $\widehat{\Phi}_{\upsilon,\eta}$, that is, $\widehat{V} \in \arg\min_{V \in \mathrm{St}(d,\ell)} \widehat{\Phi}_{\upsilon,\eta}(V)$. Then, for any $\delta \in (0, 1)$, the following inequality holds with probability at least $1 - \delta$:

$$\Phi_{\upsilon,\eta}(\widehat{V}) - \min_{V \in \mathrm{St}(d,\ell)} \Phi_{\upsilon,\eta}(V) \leq 9(M + \frac{M^3}{\upsilon})^2 \sqrt{\frac{\log\frac{3}{\delta}}{n}} + \mathsf{C}_{M,\Sigma_\mathcal{T},\Sigma_\mathcal{S}} \cdot \eta\ell \left(\frac{\log\frac{6d}{\delta}}{n} + \sqrt{\frac{\log\frac{6d}{\delta}}{n}} \right) \,,$$

where $\mathsf{C}_{M,\Sigma_\mathcal{T},\Sigma_\mathcal{S}}$ is a constant depending on M, $\Sigma_\mathcal{T}$, and $\Sigma_\mathcal{S}$.

The idea of the proof of Theorem 7.4.5 is to decompose the deviation $|\Phi_{\upsilon,\eta} - \widehat{\Phi}_{\upsilon,\eta}|$ into the deviation of the empirical risk from the population risk and the deviation of the empirical penalty term from the population penalty term. The first term on the right-hand side of the inequality of Theorem 7.4.5 corresponds to the bound on the deviation of the risk, while the other term bounds the deviation of the penalty term. The complete proof of Theorem 7.4.5 is provided in Appendix C in the Supplementary Material together with the exact form of the constant $\mathsf{C}_{M,\Sigma_\mathcal{T},\Sigma_\mathcal{S}}$.

7.5 Real-World Data Examples

Having established a framework for the first-order optimization of (7.3.6), we apply the method to several real world data sets exhibiting distribution shifts. The following examples naturally admit distinct environments with markedly different covariate distributions. We aim to verify that the causal structure we propose captures real-world relationships in data, and empirically explore the stability risk trade-off motivated in the preceding section.

7.5.1 Revisiting the Motivating Example of Section 7.1.2

We begin by revisiting the example from the introduction. In his work, Tabellini argues *Economic Output* in the 1990s is causally related to various measures of social and economic prosperity. These complex causal relationships lead to a rich data setting for domain adaptation. We consider two environments—historically rural areas (source) and urban areas (target)—and aim to create a linear predictive model for *Economic Output*. Tabellini's work suggests both covariate and concept shifts may be present. For example, historic *Urbanization* levels likely impacted other features (e.g., 'Culture' in the 1990s), and the response variable, *Economic Output*.

A simple regression may be problematic if we want a linear predictor that performs well in both rural and urban areas (Proposition 7.2.2). The issue is exacerbated when we lack equal access to information across the urbanization spectrum. Suppose we have access to rural data (*Urbanization* $< 2\%$), but only limited labeled data from urban regions (*Urbanization* $> 10\%$). We can cast this problem in the framework of Figure 7.1—with X and Y as below, while we take *Urbanization* as an environment variable (a component of E):

$$X = \{\textit{Institutions}, \textit{Literacy}, \textit{Enrollment}, \textit{Culture}, \textit{Urbanization}\},\ Y = \{\textit{Economic Output}\}\ .$$

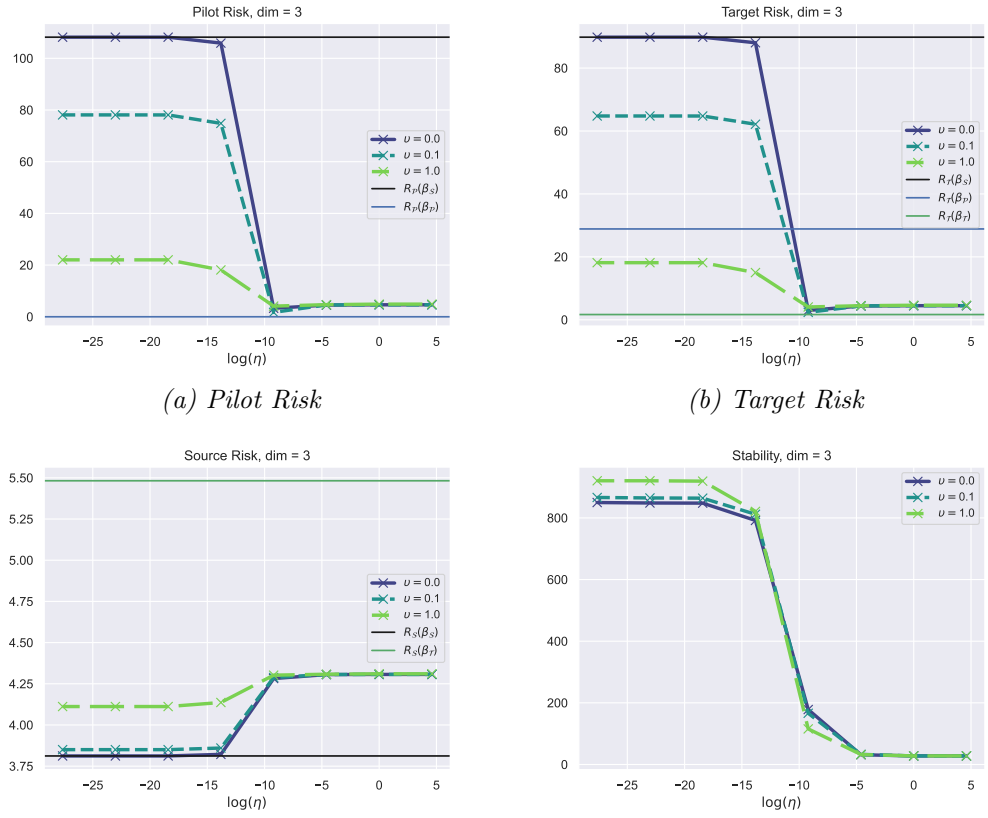
Borrowing from Tabellini's work, we would expect indicators of historical *Institutions* (pre 1850), *Literacy* (1880s), and *Enrollment* rates (1960s) to be minimally affected by *Urbanization* levels in the 1850s. Hence, we postulate the existence of a lower-dimensional variable Z with $k = 3$ that is invariant to the shift of *Urbanization*.

As our goal is prediction, not identification, we do not assume knowledge of any causal structure beyond $X \to Y$. Particularly, we do not suppose *Institutions* act as a valid instrument, and do not preempt its role as Z. Instead, we run our method, obtaining a linear subspace predictor $V\alpha_V$ by solving (7.3.6) for different regularization parameters υ, η and Stiefel dimension ℓ. For hyperparameter selection, and in line with our data constraints,

we use a small amount of pilot data for validation ($\approx 20\%$ of target data is labeled).

We find that $\ell = 3$ is an optimal lower-dimensional projection through pilot data validation. Figure 7.2 illustrates the performance of the obtained predictor with this specification. Figure 7.2 consists of 4 subplots: (a) the pilot risk, accuracy of the obtained predictor $V\alpha_V$ on a minuscule amount of labeled target data, (b) the target risk $R_{\mathcal{T}}(V\alpha_V)$, the primary objective that is inaccessible, (c) the source risk $R_{\mathcal{S}}(V\alpha_V)$, and (d) $\|V^\top(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})V\|_{\mathrm{F}}$, a quantifier for the stability/invariance.

While we have assumed no knowledge of the underlying causal structure, by navigating the stability-risk tradeoff, we find synergy with an instrumental variable perspective. With



(a) Pilot Risk *(b) Target Risk*

(c) Source Risk *(d) Stability*

Figure 7.2: Performance of the linear subspace predictor $V\alpha_V$ obtained by solving (7.3.6) for different values of (υ, η) with $d = 5$ and $\ell = 3$. The subplots (a)-(c) plot the risk of the estimator on Pilot, Target, and Source data, respectively. (d) displays $\|V^\top(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})V\|_{\mathrm{F}}$, a quantifier for the stability. In (a)-(c), the black line tracks the risk of $\beta_{\mathcal{S}}$, the blue line tracks Pilot Data OLS, and the green line tracks the risk of $\beta_{\mathcal{T}}$—a best possible oracle benchmark infeasible in practice.

a larger stability parameter η, the lower dimensional subspace prioritizes *Institutions* as a stable feature. This is emphasized in Figure 7.3, where we optimize for a 1-dimensional subspace and track the direction of the found projection with respect to each feature. While we do not claim that maximizing for stability recovers instrumental features, we certainly expect features unaffected by environmental influence to be prominent in the dominant eigen-directions of V as $\eta \to \infty$. For this example, maximizing stability improves target predictive power. We will see this is not always the case. In general, aiming for target prediction involves navigating a trade-off between source prediction and stability.

7.5.2 ML Predictions under Distribution Shifts

Next, we apply our method to three real-world datasets with less transparent causal relationships. These examples represent typical ML prediction challenges and provide a valuable benchmark for testing the robustness of our method in the absence of rigorously justified SEMs. Additionally, we introduce a data-driven heuristic for hyperparameter selection that does not depend on pilot data.

- **Forest Fires Data** [Cortez and Morais, 2007]: The goal is to predict the burned

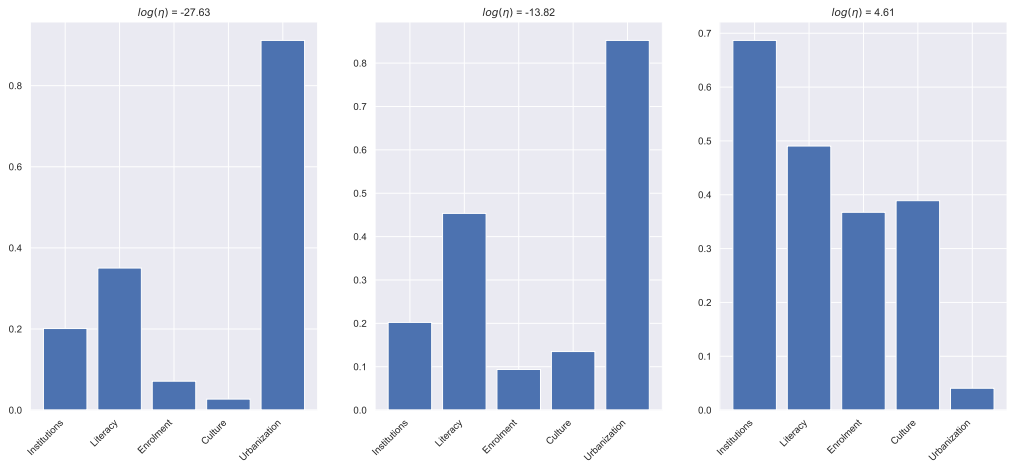


Figure 7.3: Results for $v = 0$ and $\ell = 1$ with $\log \eta \in \{-27.63, -13.82, 4.61\}$. We track the weighting placed on each feature by plotting the absolute value of each coefficient in the found $V \in \mathbb{R}^{d \times 1}$. As we increase η (from left to right), greater emphasis is placed on Institutions, the valid instrument.

area of forest fires in the northeast region of Portugal using various meteorological features. We choose seven covariates ($d = 7$): temperature, relative humidity, wind speed, precipitation, and three fire weather indices. We emulate seasonal shifts by considering two environments separated in time; the data corresponding to June, July, and August constitute the target, while we take the remaining data as the source.

- **Bike Sharing Data** [Fanaee-T and Gama, 2014]: Obtained from the bikeshare system called Capital Bikeshare serving Washington, D.C., USA, the goal of this data set is to predict hourly counts of bike rentals using features representing weather information. We postulate two different environments depending on the seasons, spring and fall for source and target, respectively. Here, we take a subset of this data by focusing on work-days and take six features ($d = 6$): hour (0 to 23), temperature, feeling temperature, humidity, wind speed, and an indicator for working day.
- **Wine Quality Data** [Cortez et al., 2009]: This exercise aims to predict the quality of wine using eleven physicochemical features ($d = 11$): fixed acidity, volatile acidity, citric acid, residual sugar, chlorides, free sulfur dioxide, total sulfur dioxide, density, pH, sulphates, and alcohol. We use the data corresponding to white and red wine as the source and target, respectively.

Unlike the example in Section 7.5.1, there are no well-studied models or instruments for these data sets. Accordingly, there is no clear choice for the hyperparameters: ℓ (Stiefel dimension), η (stability parameter) and υ (ridge parameter). It is recommended to use a small amount of pilot data for validation if available. However, when pilot data is unavailable and we have no prior knowledge of the causal structure, it is inevitable to invoke a certain rule of thumb for hyperparameter selection.

Data-Driven Hyperparameter Selection For such cases, we provide the following guidelines for hyperparameter selection. First, a principled way to choose the Stiefel dimension ℓ is to use principal component analysis (PCA). To find a stable subspace, it is reasonable

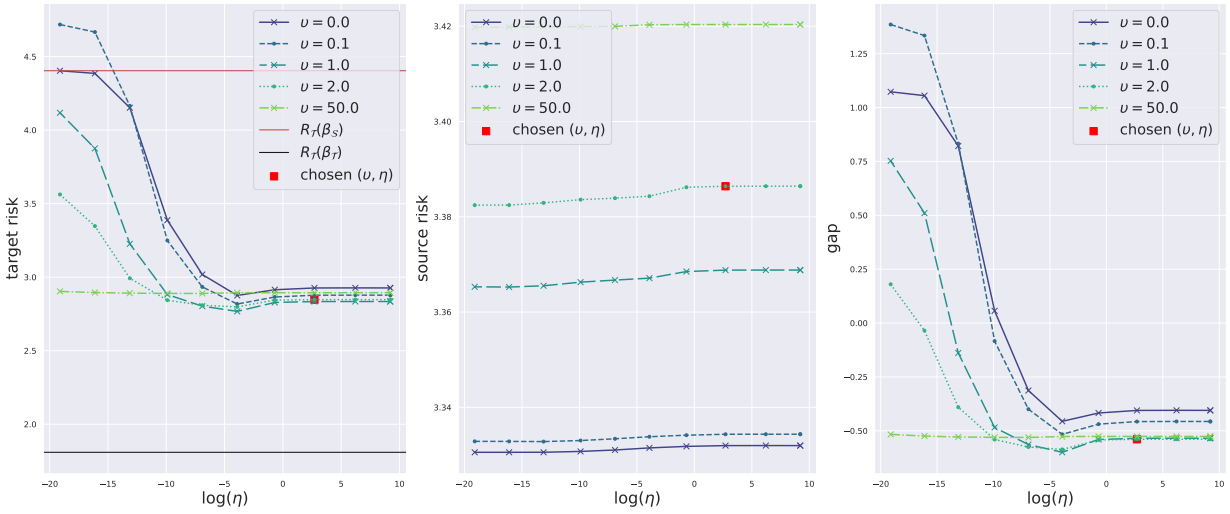
to apply PCA to the pooled covariates from both source and target and choose the number of principal components that explain most of the variance; concretely, we recommend the smallest ℓ such that the cumulative explained variance ratio exceeds a certain threshold, say 0.9. For the regularization parameters η, υ, we recommend scaling them by balancing the three terms in (7.3.4). For the ridge parameter υ, we first standardize the source covariates and then choose υ based on the ratio between the minimum source risk and the squared norm of the source risk minimizer, namely, $\frac{R_{\mathcal{S}}(\beta_{\mathcal{S}})}{\|\beta_{\mathcal{S}}\|^2}$. While letting $\upsilon \approx \frac{R_{\mathcal{S}}(\beta_{\mathcal{S}})}{\|\beta_{\mathcal{S}}\|^2}$ balances the source risk and the regularization term, we recommend a smaller value to avoid excessive regularization, say $\upsilon \approx \frac{0.1 \cdot R_{\mathcal{S}}(\beta_{\mathcal{S}})}{\|\beta_{\mathcal{S}}\|^2}$. For the stability parameter η, we first find $V \in \mathrm{St}(d, \ell)$ that minimizes (7.3.4) without the prediction term, namely, $\min_{V \in \mathrm{St}(d,\ell)} \|V^\top (\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}}) V\|_{\mathrm{F}}^2$. Then, we choose η by balancing the obtained stability term and the minimum source risk, namely, $\eta \approx \frac{2R_{\mathcal{S}}(\beta_{\mathcal{S}})}{\|V^\top (\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}}) V\|_{\mathrm{F}}^2}$. These data-driven guidelines, albeit heuristic, produce robust empirical performances, as shown below.

Results Figures 7.4, 7.5, and 7.6 illustrate the performance of the obtained predictor for the forest fires, bike sharing, and wine quality data, respectively. Each figure consists of three subfigures: (a) the target risk $R_{\mathcal{T}}(V\alpha_V)$, the primary objective that is inaccessible, (b) the source risk $R_{\mathcal{S}}(V\alpha_V)$, and (c) the difference $R_{\mathcal{T}}(V\alpha_V) - R_{\mathcal{S}}(V\alpha_V)$, a quantifier for the stability/invariance. For (a), the solid red horizontal line shows $R_{\mathcal{T}}(\beta_{\mathcal{S}})$, the target risk of the vanilla source risk minimizer $\beta_{\mathcal{S}}$, and the solid black line shows $R_{\mathcal{T}}(\beta_{\mathcal{T}})$, the target risk of the target risk minimizer $\beta_{\mathcal{T}}$ if we were given the labeled access to the target distribution $\mathcal{P}_{\mathcal{T}}(X, Y)$—a best possible oracle benchmark infeasible in practice.

From Figures 7.4(a), 7.5(a), and 7.6(a), we can see that for each υ, there is a range of η such that the resulting target risk, $R_{\mathcal{T}}(V\alpha_V)$, is smaller than the target risk of the source risk minimizer $R_{\mathcal{T}}(\beta_{\mathcal{S}})$ (visually below the red solid line). This indicates that the proposed procedure can improve the target risk by balancing the stability and the predictive power of the estimator for certain combinations of the parameters υ and η. Notably, for the wine

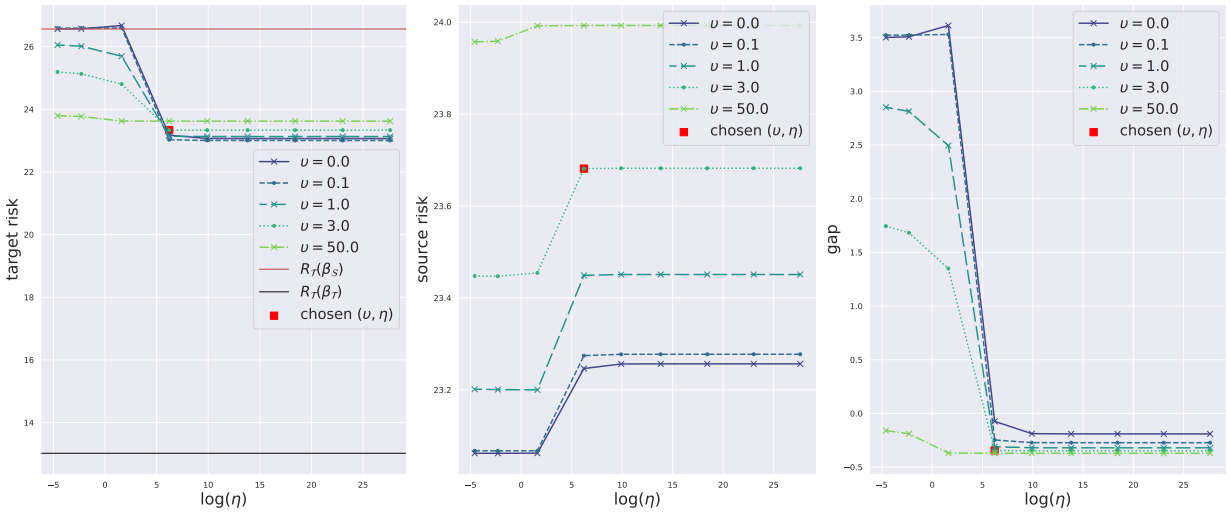
quality data, we can see improvement over the source risk minimizer $\beta_{\mathcal{S}}$ for any pair (υ, η), where we also achieve significant dimension reduction $\frac{\ell}{d} \approx 0.64$. Meanwhile, for the forest fires and bike sharing data, certain combinations of (υ, η) did not improve the target risk over the source risk. Based on the aforementioned hyperparameter selection guidelines, we obtain $(\upsilon, \eta) \approx (2, 15)$ for the forest fires data, $(\upsilon, \eta) \approx (3, 500)$ for the bike sharing data, and $(\upsilon, \eta) \approx (7, 1000)$ for the wine quality data. For these choices of hyperparameters, we find that the target risk is improved over the source risk for all data sets.

Meanwhile, we observe the non-monotonic relationship between target risk and η for fixed υ, which is more pronounced in Figures 7.4(a) and 7.6(a). This U-shaped behavior as a function of η results from the conflicting monotone behaviors of source risk and the risk gap: source risks shown in Figures 7.4(b) and 7.6(c) monotonically increase as η increases, while the risk gap shown in the corresponding (c) is almost monotone decreasing, which is explored in Theorem 7.4.3. Figures 7.4(c), 7.5(c), and 7.6(c) numerically validate the stability bound in Theorem 7.4.3: as η increases, the risk gap tends to decrease in most cases.



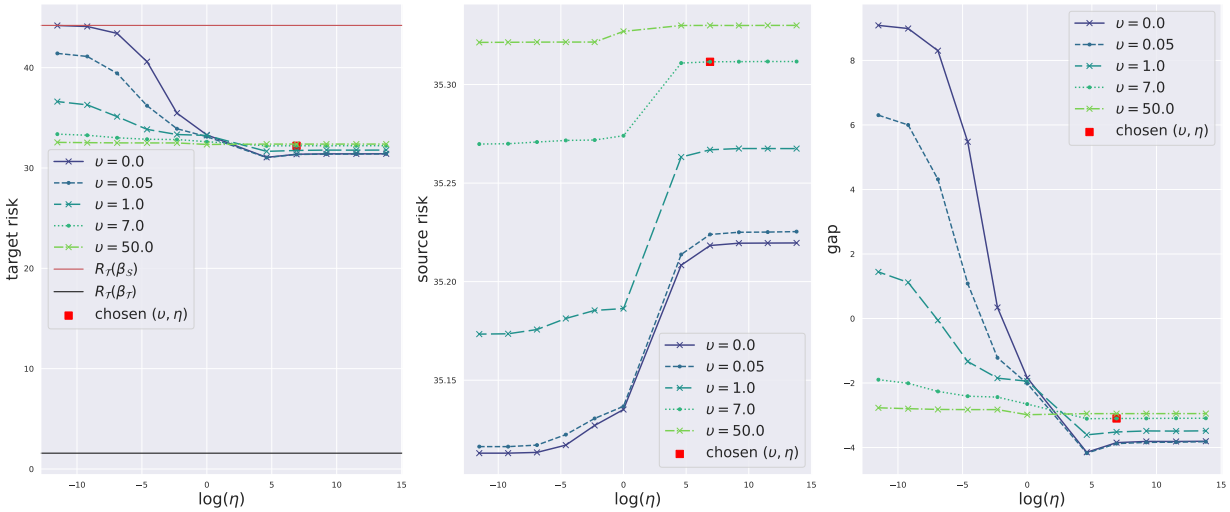
(a) Target Risk *(b) Source Risk* *(c) Target Risk - Source Risk*

Figure 7.4: Forest Fires Data: Performance of the linear subspace predictor $V\alpha_V$ obtained by solving (7.3.6) for different values of (υ, η) with $d = 7$ and $\ell = 5$. (a) plots the risk on target dataset $R_{\mathcal{T}}(V\alpha_V)$, where the solid red horizontal line shows $R_{\mathcal{T}}(\beta_{\mathcal{S}})$ and the solid black line shows $R_{\mathcal{T}}(\beta_{\mathcal{T}})$. (b) shows the risk on the source dataset $R_{\mathcal{S}}(V\alpha_V)$. (c) plots the difference $R_{\mathcal{T}}(V\alpha_V) - R_{\mathcal{S}}(V\alpha_V)$.



(a) Target Risk *(b) Source Risk* *(c) Target Risk - Source Risk*

Figure 7.5: Bike Sharing Data: Performance of the linear subspace predictor $V\alpha_V$ obtained by solving (7.3.6) with $d = 6$ and $\ell = 5$. The subplots (a)-(c) plot the same quantities as in Figure 7.4.



(a) Target Risk *(b) Source Risk* *(c) Target Risk - Source Risk*

Figure 7.6: Wine Quality Data: Performance of the linear subspace predictor $V\alpha_V$ obtained by solving (7.3.6) with $d = 11$ and $\ell = 7$. The subplots (a)-(c) plot the same quantities as in Figure 7.4.

7.6 Conclusion

This chapter investigates domain adaptation for observational data while addressing the challenge of unobserved confounding. Unobserved confounding complicates domain adaptation by (i) altering the optimal statistical models across environments and (ii) extrapolating the model to the unseen domain of covariates. To address these challenges, we propose a causal

model for distribution shifts in the presence of confounding. We highlight the limitations of traditional source risk minimization and the advantages of using an invariant subspace to stabilize learning and reduce target risk.

Methodologically, we introduce a domain adaptation algorithm that learns a representation through a lower-dimensional projection and tackles a non-convex manifold optimization problem using Riemannian optimization techniques. This algorithm acts as a stability-regularized source risk minimization that aligns with existing literature on stability and predictability in domain adaptation. Theoretically, we analyze the non-convex manifold optimization landscape and establish guarantees for nearly all local optima. With proper regularization, local optima from first-order Riemannian optimization will converge to an invariant linear subspace resistant to concept and covariate shifts. We also prove a target risk bound, showing that a predictive model derived from the learned subspace can achieve a nearly ideal gap between target and source risks.

The low-dimensional projection our method seeks will implicitly balance invariance and source domain predictive power to minimize target risk. This will be governed by hyperparameters (η, υ) and Stiefel dimension ℓ. Therefore, in practice, the choice of these hyperparameters is crucial. Hyperparameter selection guidelines in Section 7.5.2 provide actionable heuristics, which we verify to be effective from the examples in Section 7.5, but we believe more systematic hyperparameter selection rules could be more helpful. We leave this for future work. The current method is based on linear projections and linear regression; see Appendix C in the Supplementary Material for further discussions on the testability of the main assumption and comparison to the usual instrumental variable setting. It would be interesting to extend the proposed framework to partially linear or nonlinear models for full generality, as well as to the setting with multiple source environments. We leave these as future directions.

In summary, our method is a data-driven approach that extends the utility of instrumental variables for identification problems to domain adaptation by automatically finding

the stable representation. Amid increasing interest in ML/AI, our method shows how to utilize traditional statistical/econometric techniques for data discovery in the AI context by properly designing automated procedures.

REFERENCES

P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Brian DO Anderson. Reverse-time diffusion equation models. *Stochastic Processes and their Applications*, 12(3):313–326, 1982.

Joshua D. Angrist and Guido W. Imbens. Two-stage least squares estimation of average causal effects in models with variable treatment intensity. *Journal of the American Statistical Association*, 90(430):431–442, 1995.

Martin Arjovsky, Léon Bottou, Ishaan Gulrajani, and David Lopez-Paz. Invariant risk minimization. *arXiv preprint arXiv:1907.02893*, 2019.

Jacob Austin, Daniel D Johnson, Jonathan Ho, Daniel Tarlow, and Rianne Van Den Berg. Structured denoising diffusion models in discrete state-spaces. *Advances in Neural Information Processing Systems*, 34:17981–17993, 2021.

Dale Belman and John S. Heywood. Sheepskin effects in the returns to education: An examination of women and minorities. *The Review of Economics and Statistics*, 73(4): 720–724, 1991.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. In *Advances in Neural Information Processing Systems*, 2006.

Shai Ben-David, John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman Vaughan. A theory of learning from different domains. *Machine Learning*, 79:151–175, 2010.

Joe Benton, Valentin De Bortoli, Arnaud Doucet, and George Deligiannidis. Nearly d-linear convergence bounds for diffusion models via stochastic localization. *arXiv preprint arXiv:2308.03686*, 2023.

John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128. Association for Computational Linguistics, 2006.

John Blitzer, Koby Crammer, Alex Kulesza, Fernando Pereira, and Jennifer Wortman. Learning bounds for domain adaptation. In *Advances in Neural Information Processing Systems*, 2007.

Adam Block, Youssef Mroueh, and Alexander Rakhlin. Generative modeling with denoising auto-encoders and langevin sampling. *arXiv preprint arXiv:2002.00107*, 2020.

Richard Blundell, Xiaohong Chen, and Dennis Kristensen. Semi-nonparametric IV estimation of shape-invariant engel curves. *Econometrica*, 75(6):1613–1669, 2007. doi:10.1111/j.1468-0262.2007.00808.x.

Yann Brenier. Décomposition polaire et réarrangement monotone des champs de vecteurs. *CR Acad. Sci. Paris Sér. I Math.*, 305:805–808, 1987.

Stefano Bruno, Ying Zhang, Dong-Young Lim, Ömer Deniz Akyildiz, and Sotirios Sabanis. On diffusion-based generative models and their error bounds: The log-concave case with full convergence estimates. *arXiv preprint arXiv:2311.13584*, 2023.

David Card. Using geographic variation in college proximity to estimate the return to schooling. In Louis N. Christofides, E. Kenneth Grant, and Robert Swidinsky, editors, *Aspects of Labour Market Behaviour: Essays in Honour of John Vanderkamp*, pages 201–222. University of Toronto Press, Toronto, 1995.

David Card. Estimating the return to schooling: Progress on some persistent econometric problems. *Econometrica*, 69(5):1127–1160, 2001. doi:10.1111/1468-0262.00237.

David Card and Alan B. Krueger. Minimum wages and employment: A case study of the fast-food industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4): 772–793, 1994.

Pedro Carneiro, James J. Heckman, and Edward J. Vytlacil. Estimating marginal returns to education. *American Economic Review*, 101(6):2754—2781, October 2011. doi:10.1257/aer.101.6.2754.

Raymond J. Carroll, David Ruppert, Leonard A. Stefanski, and Ciprian M. Crainiceanu. *Measurement Error in Nonlinear Models: A Modern Perspective*. Chapman and Hall/CRC, Boca Raton, FL, 2 edition, 2006. doi:10.1201/9781420010138.

Gary Chamberlain. Asymptotic efficiency in estimation with conditional moment restrictions. *Journal of Econometrics*, 34(3):305–334, 1987. ISSN 0304-4076. doi:10.1016/0304-4076(87)90015-7.

Hongrui Chen, Holden Lee, and Jianfeng Lu. Improved analysis of score-based generative modeling: user-friendly bounds under minimal smoothness assumptions. In *International Conference on Machine Learning*, pages 4735–4763. PMLR, 2023.

Sitan Chen, Sinho Chewi, Jerry Li, Yuanzhi Li, Adil Salim, and Anru R Zhang. Sampling is as easy as learning the score: theory for diffusion models with minimal data assumptions. *arXiv preprint arXiv:2209.11215*, 2022.

Sitan Chen, Sinho Chewi, Holden Lee, Yuanzhi Li, Jianfeng Lu, and Adil Salim. The probability flow ode is provably fast. *Advances in Neural Information Processing Systems*, 36, 2024.

Xiaohong Chen and Demian Pouzo. Estimation of nonparametric conditional moment models with possibly nonsmooth generalized residuals. *Econometrica*, 80(1):277–321, 2012. doi:10.3982/ECTA7888.

Rune Christiansen, Niklas Pfister, Martin Emil Jakobsen, Nicola Gnecco, and Jonas Peters. A causal framework for distribution generalization. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):6614–6630, 2022.

Paulo Cortez and Aníbal Morais. A data mining approach to predict forest fires using meteorological data. In *Portuguese Conference on Artificial Intelligence*, 2007.

Paulo Cortez, António Cerdeira, Fernando Almeida, Telmo Matos, and José Reis. Modeling wine preferences by data mining from physicochemical properties. *Decision Support Systems*, 47(4):547–553, 2009.

Serge Darolles, Yanqin Fan, Jean-Pierre Florens, and Eric Renault. Nonparametric instrumental regression. *Econometrica*, 79(5):1541–1565, 2011. doi:10.3982/ECTA6539.

Nabarun Deb and Tengyuan Liang. No-regret generative modeling via parabolic monge-ampere pde, 2026.

Valentin DeBortoli, James Thornton, Jeremy Heng, and Arnaud Doucet. Diffusion schrödinger bridge with applications to score-based generative modeling. *Advances in Neural Information Processing Systems*, 34:17695–17709, 2021.

Aurore Delaigle and Irène Gijbels. Practical bandwidth selection in deconvolution kernel density estimation. *Computational Statistics & Data Analysis*, 45(2):249–267, 2004. doi:10.1016/S0167-9473(02)00329-8.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Kulunu Dharmakeerthi, Yousef El-Laham, Henry H. Wong, Vamsi K. Potluru, Changhong X. He, and Taosong He. Beyond linear diffusions: Improved representations for rare conditional generative modeling. In *NeurIPS 2025 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2025. URL `https://openreview.net/forum?id=dGtlMztVHU`.

Kulunu Dharmakeerthi, YoonHaeng Hur, and Tengyuan Liang. Learning when the concept shifts: Confounding, invariance, and dimension reduction. *Journal of the American Statistical Association*, 0(ja):1–24, 2026. doi:10.1080/01621459.2026.2656457.

Nishanth Dikkala, Greg Lewis, Lester Mackey, and Vasilis Syrgkanis. Minimax estimation of conditional moment models. In *Advances in Neural Information Processing Systems*, volume 33, 2020.

Tim Dockhorn, Arash Vahdat, and Karsten Kreis. Score-based generative modeling with critically-damped langevin diffusion. *arXiv preprint arXiv:2112.07068*, 2021.

Bradley Efron and Carl Morris. Stein's estimation rule and its competitors—an empirical bayes approach. *Journal of the American Statistical Association*, 68(341):117–130, 1973.

Jianqing Fan. On the optimal rates of convergence for nonparametric deconvolution problems. *The Annals of Statistics*, 19(3):1257–1272, 1991. doi:10.1214/aos/1176348248.

Hadi Fanaee-T and Joao Gama. Event labeling combining ensemble detectors and background knowledge. *Progress in Artificial Intelligence*, 2:113–127, 2014.

Joachim Freyberger. On completeness and consistency in nonparametric instrumental variable models. *Econometrica*, 85(5):1629–1644, 2017. doi:10.3982/ECTA13304.

Yaroslav Ganin, Evgeniya Ustinova, Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, Mario March, and Victor Lempitsky. Domain-adversarial training of neural networks. *Journal of Machine Learning Research*, 17(59):1–35, 2016.

Xuefeng Gao, Hoang M Nguyen, and Lingjiong Zhu. Wasserstein convergence guarantees for a general class of score-based generative models. *Journal of machine learning research*, 26 (43):1–54, 2025.

Jiawei Ge, Shange Tang, Jianqing Fan, Cong Ma, and Chi Jin. Maximum likelihood estimation is all you need for well-specified covariate shift. *arXiv preprint arXiv:2311.15961*, 2023.

Marta Gentiloni-Silveri and Antonio Ocello. Beyond log-concavity and score regularity: Improved convergence bounds for score-based generative models in w2-distance. *arXiv preprint arXiv:2501.02298*, 2025.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

Christian Gourieroux, Alain Monfort, and Eric Renault. Indirect inference. *Journal of Applied Econometrics*, 8(S1):S85–S118, 1993.

Wenxuan Guo, YoonHaeng Hur, Tengyuan Liang, and Chris Ryan. Online learning to transport via the minimal selection principle. In Po-Ling Loh and Maxim Raginsky, editors, *Proceedings of Thirty Fifth Conference on Learning Theory*, volume 178 of *Proceedings of Machine Learning Research*, pages 4085–4109. PMLR, 02–05 Jul 2022.

Peter Hall and Joel L. Horowitz. Nonparametric methods for inference in the presence of instrumental variables. *The Annals of Statistics*, 33(6):2904–2929, 2005. doi:10.1214/009053605000000714.

Lars Peter Hansen. Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054, 1982. doi:10.2307/1912775.

Lars Peter Hansen and Kenneth J. Singleton. Generalized instrumental variables estimation of nonlinear rational expectations models. *Econometrica*, 50(5):1269–1286, 1982. doi:10.2307/1911873.

Jason Hartford, Greg Lewis, Kevin Leyton-Brown, and Matt Taddy. Deep IV: A flexible approach for counterfactual prediction. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1414–1423, 2017.

James J. Heckman and Edward Vytlacil. Structural equations, treatment effects, and econometric policy evaluation. *Econometrica*, 73(3):669–738, 2005. doi:10.1111/j.1468-0262.2005.00594.x.

Christina Heinze-Deml, J. Peters, and Nicolai Meinshausen. Invariant causal prediction for nonlinear models. *Journal of Causal Inference*, 6(2):20170016, 2017.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33:6840–6851, 2020.

Joel L. Horowitz. Applied nonparametric instrumental variables estimation. *Econometrica*, 79(2):347–394, 2011. doi:10.3982/ECTA8662.

Rongjie Huang, Jiawei Huang, Dongchao Yang, Yi Ren, Luping Liu, Mingze Li, Zhenhui Ye, Jinglin Liu, Xiang Yin, and Zhou Zhao. Make-an-audio: Text-to-audio generation with prompt-enhanced diffusion models. In *International Conference on Machine Learning*, pages 13916–13932. PMLR, 2023.

Thomas Hungerford and Gary Solon. Sheepskin effects in the returns to education. *The Review of Economics and Statistics*, 69(1):175–177, 1987. doi:10.2307/1937919.

YoonHaeng Hur and Tengyuan Liang. A convexified matching approach to imputation and individualized inference. *arXiv preprint arXiv:2407.05372*, 2024. doi:10.48550/arXiv.2407.05372.

YoonHaeng Hur and Tengyuan Liang. Detecting weak distribution shifts via displacement interpolation. *Journal of Business & Economic Statistics*, 43(1):178–190, 2025. doi:10.1080/07350015.2024.2335957.

YoonHaeng Hur, Wenxuan Guo, and Tengyuan Liang. Reversible gromov–monge sampler for simulation-based inference. *SIAM Journal on Mathematics of Data Science*, 6(2):283–310, 2024. doi:10.1137/23M1550384.

Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

Guido W. Imbens and Joshua D. Angrist. Identification and estimation of local average treatment effects. *Econometrica*, 62(2):467–475, 1994. doi:10.2307/2951620.

David A. Jaeger and Marianne E. Page. Degrees matter: New evidence on sheepskin effects in the returns to education. *The Review of Economics and Statistics*, 78(4):733–740, 1996.

W. James and C. Stein. Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability*, 1:361–379, 1961.

Richard Jordan, David Kinderlehrer, and Felix Otto. The variational formulation of the fokker–planck equation. *SIAM journal on mathematical analysis*, 29(1):1–17, 1998.

Mohammadreza Mousavi Kalan, Zalan Fabian, Salman Avestimehr, and Mahdi Soltanolkotabi. Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. In *Advances in Neural Information Processing Systems*, 2020.

Thomas J. Kane and Cecilia Elena Rouse. Labor-market returns to two- and four-year college. *American Economic Review*, 85(3):600–614, 1995.

Tero Karras, Miika Aittala, Timo Aila, and Samuli Laine. Elucidating the design space of diffusion-based generative models. *Advances in Neural Information Processing Systems*, 35:26565–26577, 2022.

Diederik Kingma, Tim Salimans, Ben Poole, and Jonathan Ho. Variational diffusion models. *Advances in Neural Information Processing Systems*, 34:21696–21707, 2021.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence for score-based generative modeling with polynomial complexity. *Advances in Neural Information Processing Systems*, 35: 22870–22882, 2022.

Holden Lee, Jianfeng Lu, and Yixin Tan. Convergence of score-based generative modeling for general data distributions. In *International Conference on Algorithmic Learning Theory*, pages 946–985. PMLR, 2023.

Qi Lei, Wei Hu, and Jason Lee. Near-optimal linear regression under distribution shift. In *International Conference on Machine Learning*, 2021.

Tengyuan Liang. How well generative adversarial networks learn distributions. *Journal of Machine Learning Research*, 22(228):1–41, 2021. URL `http://jmlr.org/papers/v22/20-911.html`.

Tengyuan Liang. Blessings and curses of covariate shifts: Adversarial learning dynamics, directional convergence, and equilibria. *Journal of Machine Learning Research*, 25(140): 1–27, 2024.

Tengyuan Liang. Distributional shrinkage i: Universal denoisers in multi-dimensions. *arXiv preprint arXiv:2511.09500*, 2025a.

Tengyuan Liang. Distributional shrinkage ii: Optimal transport denoisers with higher-order scores. *arXiv preprint arXiv:2512.09295*, 2025b.

Tengyuan Liang, Kulunu Dharmakeerthi, and Takuya Koriyama. Denoising diffusions with optimal transport: Localization, curvature, and multi-scale complexity. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL `https://openreview.net/forum?id=sj1wU6gBXH`. J2C Certification.

Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, 2015.

Frederike Lübeck, Charlotte Bunne, Gabriele Gut, Jacobo Sarabia del Castillo, Lucas Pelkmans, and David Alvarez-Melis. Neural unbalanced optimal transport via cycle-consistent semi-couplings. *arXiv preprint arXiv:2209.15621*, 2022.

Sara Magliacane, Thijs van Ommen, Tom Claassen, Stephan Bongers, Philip Versteeg, and Joris M Mooij. Domain adaptation by using causal inference to predict invariant conditional distributions. In *Advances in Neural Information Processing Systems*, 2018.

Yishay Mansour, Mehryar Mohri, and Afshin Rostamizadeh. Domain adaptation: Learning bounds and algorithms. *arXiv preprint arXiv:0902.3430*, 2009.

Dimitra Maoutsa, Sebastian Reich, and Manfred Opper. Interacting particle solutions of fokker–planck equations through gradient–log–density estimation. *Entropy*, 22(8):802, 2020.

Daniel McFadden. A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57(5):995–1026, 1989.

Alexander Meister. *Deconvolution Problems in Nonparametric Statistics*, volume 193 of *Lecture Notes in Statistics*. Springer, Berlin, Heidelberg, 2009. doi:10.1007/978-3-540-87557-4.

Paul Milgrom and Ilya Segal. Envelope theorems for arbitrary choice sets. *Econometrica*, 70 (2):583–601, 2002.

Andrea Montanari. Sampling, diffusions, and stochastic localization. *arXiv preprint arXiv:2305.10690*, 2023.

Whitney K. Newey. 16 efficient estimation of models with conditional moment restrictions. In *Econometrics*, volume 11 of *Handbook of Statistics*, pages 419–454. Elsevier, 1993. doi:10.1016/S0169-7161(05)80051-3.

Whitney K. Newey and James L. Powell. Instrumental variable estimation of nonparametric models. *Econometrica*, 71(5):1565–1578, 2003. doi:10.1111/1468-0262.00459.

Alexander Quinn Nichol and Prafulla Dhariwal. Improved denoising diffusion probabilistic models. In *International Conference on Machine Learning*, pages 8162–8171. PMLR, 2021.

James R Norris. *Markov chains*. Cambridge university press, 1998.

Kazusato Oko, Shunta Akiyama, and Taiji Suzuki. Diffusion models are minimax optimal distribution estimators. In *International Conference on Machine Learning*, pages 26517–26582. PMLR, 2023.

Philip Oreopoulos and Uros Petronijevic. Making college worth it: A review of the returns to higher education. *The Future of Children*, 23(1):41–65, 2013.

Ariel Pakes and David Pollard. Simulation and the asymptotics of optimization estimators. *Econometrica*, 57(5):1027–1057, 1989.

Victor M. Panaretos. A statistician's view on deconvolution and unfolding. In *Proceedings of the PHYSTAT 2011 Workshop on Statistical Issues Related to Discovery Claims in Search Experiments and Unfolding*, pages 229–239, Geneva, 2011. CERN. URL `https://indico.cern.ch/event/107747/contributions/32644/attachments/24315/34997/panaretos.pdf`.

George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. Normalizing flows for probabilistic modeling and inference. *Journal of Machine Learning Research*, 22(57):1–64, 2021.

Judea Pearl. *Causality*. Cambridge University Press, 2 edition, 2009.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: Identification and confidence intervals. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 78(5):947–1012, 2016.

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. *Elements of Causal inference: Foundations and Learning Algorithms*. The MIT Press, 2017.

Niklas Pfister, Peter Bühlmann, and Jonas Peters. Invariant causal prediction for sequential data. *Journal of the American Statistical Association*, 114(527):1264–1276, 2019.

Herbert Robbins. An empirical bayes approach to statistics. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 157–163, Berkeley, CA, 1956. University of California Press.

Herbert Robbins. The empirical bayes approach to statistical decision problems. *The Annals of Mathematical Statistics*, 35(1):1–20, 1964.

Herbert E. Robbins. An empirical bayes approach to statistics. In *Breakthroughs in Statistics: Foundations and Basic Theory*, pages 388–394. Springer New York, 1992.

Mateo Rojas-Carulla, Bernhard Schölkopf, Richard Turner, and Jonas Peters. Invariant models for causal transfer learning. *Journal of Machine Learning Research*, 19(36):1–34, 2018.

Dominik Rothenhäusler, Nicolai Meinshausen, Peter Bühlmann, and Jonas Peters. Anchor regression: Heterogeneous data meet causality. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 83(2):215–246, 2021.

Saeed Saremi and Aapo Hyvärinen. Neural empirical bayes. *Journal of Machine Learning Research*, 20(181):1–23, 2019.

J. D. Sargan. The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415, 1958. doi:10.2307/1907619.

Bernhard Scholkopf, Dominik Janzing, J. Peters, Eleni Sgouritsa, Kun Zhang, and Joris M. Mooij. On causal and anticausal learning. In *International Conference on Machine Learning*, 2012.

Jian Shen, Yanru Qu, Weinan Zhang, and Yong Yu. Wasserstein distance guided representation learning for domain adaptation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2018.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of Statistical Planning and Inference*, 90:227–244, 2000.

Rahul Singh, Manfred Sahani, and Arthur Gretton. Kernel instrumental variable regression. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International Conference on Machine Learning*, pages 2256–2265. PMLR, 2015.

Yang Song and Stefano Ermon. Generative modeling by estimating gradients of the data distribution. *Advances in Neural Information Processing Systems*, 32, 2019.

Yang Song and Stefano Ermon. Improved techniques for training score-based generative models. *Advances in Neural Information Processing Systems*, 33:12438–12448, 2020.

Yang Song, Jascha Sohl-Dickstein, Diederik P Kingma, Abhishek Kumar, Stefano Ermon, and Ben Poole. Score-based generative modeling through stochastic differential equations. *arXiv preprint arXiv:2011.13456*, 2020.

Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum likelihood training of score-based diffusion models. *Advances in Neural Information Processing Systems*, 34: 1415–1428, 2021.

James H. Stock and Francesco Trebbi. Retrospectives: Who invented instrumental variable regression? *Journal of Economic Perspectives*, 17(3):177–194, 2003.

Guido Tabellini. Culture and institutions: Economic development in the regions of Europe. *Journal of the European Economic Association*, 8(4):677–716, 2010.

Joel A Tropp et al. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.

Eric Tzeng, Judy Hoffman, Ning Zhang, Kate Saenko, and Trevor Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

Roman Vershynin. *High-dimensional probability: an introduction with applications in data science*. Cambridge University Press, 2018.

APPENDIX A

APPENDICES TO CHAPTER 3

A.1 Related Work

Probabilistic generative models, including generative adversarial networks, flow-based generative models, and diffusion-based generative models, have recently received broad research interest, both empirically [Goodfellow et al., 2020, Song and Ermon, 2019, 2020, Song et al., 2021, Karras et al., 2022, Nichol and Dhariwal, 2021, Kingma et al., 2021, Huang et al., 2023] and theoretically [Papamakarios et al., 2021, Liang, 2021, Hur et al., 2024, Oko et al., 2023, Lübeck et al., 2022, Chen et al., 2024, Block et al., 2020, DeBortoli et al., 2021, Lee et al., 2022, Montanari, 2023]. Generative models with diffusion-based sampling can be traced back to three influential formulations: Denoising Diffusion Probabilistic Models (DDPMs) [Ho et al., 2020, Sohl-Dickstein et al., 2015], Score-based Generative Models (SGMs) [Song and Ermon, 2019], and Score-based Stochastic Differential Equations (Score SDEs) [Song et al., 2021, Karras et al., 2022]. The latter universalizes the frameworks as discretizations of a particular Stochastic Differential Equation at specific signal-to-noise ratios (SNR). Sampling can be viewed as Langevin dynamics for the time-reversed SDE [Anderson, 1982, Ho et al., 2020, Dockhorn et al., 2021], or deterministic transports via a probability flow ODE [Song et al., 2020, Maoutsa et al., 2020, Guo et al., 2022]. The deterministic denoising outlined in Song et al. [2020] is optimal in terms of transportation of measure in the Wasserstein metric. This was previously noted by Chen et al. [2024], Jordan et al. [1998] in the infinitesimal limit $\eta \to 0$. We show in Proposition 3.2.3 that for a fixed stepsize η, the score denoising map is the optimal transport map in the backward chain.

Rigorous justification for diffuse-then-denoise can be found in the seminal work of Saremi and Hyvarinen [Saremi and Hyvärinen, 2019] who unified two distinct schemes: (1) smoothing via Gaussian convolution $\mathbf{Y} = \mathbf{X} + \frac{1}{\mathsf{r}}\mathcal{N}(0, I_d)$, and (2) denoising via the score function $\mathbb{E}[\mathbf{X}|\mathbf{Y} = y] = y + \frac{1}{\mathsf{r}^2}\nabla \log p_{\mathbf{Y}}(y)$, to design a single machine for sampling $\mathbf{X}$. Their approach

is motivated by a fundamental result in the concentration of measure literature [Vershynin, 2018]; that d-dimensional Gaussian vectors concentrate on a uniform sphere of radius $\sqrt{d}$ in high-dimensions. And so, while $\mathbf{X}$ may be non-log-concave or confined to a low-dimensional manifold, $\mathbf{Y}$ may not be. Thus, diffuse ($\mathbf{X} \to \mathbf{Y}$) then denoise ($\mathbf{Y} \to \mathbf{X}$) presents a way to sample when the distribution of $\mathbf{X}$ is misbehaved. Exploiting the machinery of measure transportation, we show that Gaussian convolution up to an appropriate signal-to-noise ratio r promotes log-concavity, yielding desirable contractive properties for denoising, $\mathbf{Y} \to \mathbf{X}$.

Recent empirical work [Song and Ermon, 2019, 2020, Song et al., 2021, Karras et al., 2022] demonstrates dramatic improvements in sample quality and log-likelihood metrics for diffusions by employing a rich scheme of noise scheduling, namely the effective SNR $t \mapsto \mathsf{r}(t)$, a map between time and the corresponding SNR. Several authors consider jointly learning the schedule alongside diffusion network parameters [Nichol and Dhariwal, 2021, Kingma et al., 2021]. While state-of-the-art applications suggest that convolving at multiple noise scales is advantageous [Dhariwal and Nichol, 2021, Austin et al., 2021], its theoretical benefit remains to be understood. Restricting to the Ornstein-Uhlenbeck process, we discover a multi-scale complexity measure that controls the effective contraction of diffuse-then-denoise at each SNR r. Particularly for non-log-concave distributions, a wide regime of SNR schedule $\mathsf{r}(t)$ may be necessary as diffusing at a specific SNR scale r pushes the problem into an effective near-log-concave setting with strong curvature. Our multi-scale complexity determines the effective curvature, thereby providing a vehicle for probing the noise-scheduling question.

Theoretical investigation of diffusion-based sampling mostly focuses on the non-expansive properties of the denoising process under f-divergence, say total variation d_{TV} or Kullback-Leibler d_{KL}, where data processing inequities hold. The current theory falls short in addressing the behavior of the backward denoising process under the Wasserstein metric, where the backward denoising chain does incur expansion even for simple Gaussian distributions [Chen et al., 2022]. Typically, the problem is coupled with the additional burden of having access to only an estimated score function [Block et al., 2020, DeBortoli et al., 2021, Lee et al., 2022].

This presents a significant challenge; however, it is not the focus of the current chapter. We isolate treatment of the backward denoising question: initializing at μ_T, and the equilibrium measure, ν, suppose we observe $\{\mu_0 \leftarrow_{\mathbf{b}_1^\mu} \dots \leftarrow_{\mathbf{b}_K^\mu} \mu_K\}$, $\{\bar{\nu}_0 \leftarrow_{\mathbf{b}_1^\mu} \dots \leftarrow_{\mathbf{b}_K^\mu} \bar{\nu}_K := \nu\}$. How does $d(\mu_K, \bar{\nu}_K) \to d(\mu_{K-1}, \bar{\nu}_{K-1}) \to \cdots \to d(\mu_0, \bar{\nu}_0)$ evolve? Is the backward denoising chain expansive or contractive at each time scale $t = k\eta$? What geometric complexities govern the quality of the denoising at each time scale? We view our focus as complementary to the current literature, which conducts careful sensitivity analyses of score estimation and time discretization.

Provided the score functions are accurate in an L_2 sense, Block et al. [2020], Lee et al. [2022] established results in the Wasserstein metric for unimodal distributions, for example, those satisfying strong dissipativity or a log-Sobolev inequality. These results are extended in Lee et al. [2023], Chen et al. [2022, 2023] to allow substantial non-log-concavity for reverse SDE sampling, and in Chen et al. [2024] for the probability flow ODE. However, these analyses are conducted in d_{TV} or d_{KL} and avoid the detailed curvature information in the backward map. For d_{TV} or d_{KL}, and a finite K, one can always appeal to a data processing inequality to bound $d(\mu_0, \bar{\mu}_0) \leq d(\mu_K, \nu)$, solely determined by the forward diffusion chain. This is not fine-grained enough to understand denoising. It overlooks the curvature information and the amount of non-log-concavity at any specific SNR r. As in the introduction, the backward denoising step could expand significantly or contract much faster than the forward diffusion under the Wasserstein metric.

More recently, Bruno et al. [2023], Gao et al. [2025] establish results in the Wasserstein metric under log-concavity assumptions. The concurrent work [Gentiloni-Silveri and Ocello, 2025] tackles the problem assuming weak log-concavity of the target density, a property with origins in dissipativity - that the target density is essentially log-concave outside a region of compact support. Our work makes no such assumptions. Similar to Gentiloni-Silveri and Ocello [2025], we note that the denoising process can exhibit non-contractive behavior. In our analysis, the fine-grained behavior of the denoising process is governed by a notion

of average curvature, defined as an integrated tail function measuring the relative mass of locations with positive curvature versus those with negative curvature. Notions of average curvature have been used to derive improved convergence bounds for diffusion models with respect to d_{KL} [Chen et al., 2023, Benton et al., 2023]. Our work shows that a different notion of average curvature also plays a role in the sensitivity analysis when the metric is the Wasserstein distance. Our perspective helps bridge the parallel theoretical analyses of diffusion models under f-divergence and Wasserstein distance.

More importantly, previous approaches disconnect denoising from diffusion when evaluating the success of score-based sampling. We show that the net benefit of the diffuse-then-denoise process matters: the contraction of the forward diffusion, offset by the possible backward expansion, results in a net benefit as long as there is enough negative curvature on average. This paradigm shift is enabled by a new notion of multi-scale complexity governing the effective curvature and "localization" effect of the diffuse-then-denoise process. This allows an analysis that extends beyond log-concavity assumptions.

A.2 Technical Proofs

The rest of the Appendix is organized as follows.

Section A.2.1 gathers the technical preparations for denoising and localization, where we prove the results in Section 3.2. We derive the optimal transport denoising map and explain how the curvature function appears from a localization perspective.

Section A.2.2 presents analyses of the forward and backward processes, where we prove the results in Section 3.3. We demonstrate how the curvature function influences backward denoising via a sensitivity analysis under the Wasserstein metric. We also analyze the diffuse-then-denoise process, showing that adding noise then denoising can be beneficial in the presence of negative curvature.

Section A.2.3 provides proofs for the results in Section 3.4. We extend beyond the log-concave setting by defining and analyzing a multi-scale complexity that captures an average

notion of curvature across all time scales. This multi-scale complexity plays a key role in the sensitivity analysis of the diffusion model under the Wasserstein metric.

A.2.1 Proofs for Section 3.2

Proof of Proposition 3.2.3. By Lemma 10.1.2 in Ambrosio et al. [2008],

$$\frac{1}{\eta}(\mathbf{t}_{\mu_{t+\eta}}^{\mu_t} - \mathbf{i}) \in \partial\mathcal{G}(\mu_{t+\eta})$$

where $\partial\mathcal{G}(\mu_{t+\eta})$ is the strong subdifferential. Recall that $\mathcal{G}(\nu) = \mathcal{F}(\nu) + \beta^{-1}\mathcal{E}(\nu)$. For a given ν with density $\nu = \rho \cdot \mathcal{L}^d$, by Lemma 10.4.1 in Ambrosio et al. [2008], we know any $\boldsymbol{\xi}(x) \in L^2(\nu; \mathbb{R}^d)$ in $\partial\mathcal{G}(\nu)$ admits the representation

$$\boldsymbol{\xi}(x) = \nabla\frac{\delta\mathcal{G}}{\delta\rho} = \nabla f(x) + \beta^{-1}\nabla\log\rho(x), \text{ for } \nu\text{-a.e. } x \in \mathbb{R}^d .$$

Take $\nu = \mu_{t+\eta}$, we finish the proof. □

Proof of Proposition 3.2.4. This result follows from Tweedie's formula [Robbins, 1992]. The conditional distribution of $\mathbf{Y} = y$ given $\mathbf{X}$ admits the following density: $p_{\mathbf{Y}}(y|\mathbf{X}) := \frac{\mathrm{d}}{\mathrm{d}y}P(\mathbf{Y} \le y|\mathbf{X}) = \frac{1}{\sigma}\phi((y - \mathbf{X})/\sigma)$ where $\phi : \mathbb{R}^d \to \mathbb{R}$ is the density of d-dimensional standard normal. This implies

$$\nabla p_{\mathbf{Y}}(y|\mathbf{X}) = -\frac{1}{\sigma^2}(y - \mathbf{X})p_{\mathbf{Y}}(y|\mathbf{X}) \quad \text{and} \quad \nabla p_{\mathbf{Y}}(y) = -\frac{1}{\sigma^2}\mathbb{E}[(y - \mathbf{X})p_{\mathbf{Y}}(y|\mathbf{X})]. \quad \text{(A.2.1)}$$

Therefore, we obtain

$$\frac{\nabla p_{\mathbf{Y}}(y)}{p_{\mathbf{Y}}(y)} = -\frac{1}{\sigma^2}\mathbb{E}[(y - \mathbf{X})\frac{p_{\mathbf{Y}}(y|\mathbf{X})}{p_{\mathbf{Y}}(y)}] = -\frac{1}{\sigma^2}\big\{y - \mathbb{E}[\mathbf{X}|\mathbf{Y} = y]\big\} ,$$

where we used the fact that $\mathbb{E}[h(\mathbf{X})\frac{p_{\mathbf{Y}}(y|\mathbf{X})}{p_{\mathbf{Y}}(y)}] = \mathbb{E}[h(\mathbf{X})|\mathbf{Y} = y]$ for any integrable function h with respect to $\mathbf{X}$. This completes the proof. □

Proof of Proposition 3.2.5. Taking the derivative of (A.2.1) with respect to y, we have

$$\nabla^2 p_{\mathbf{Y}}(y|\mathbf{X}) = \Big[-\frac{1}{\sigma^2} I_d + \frac{1}{\sigma^4}(y-\mathbf{X})(y-\mathbf{X})^\top\Big] p_{\mathbf{Y}}(y|\mathbf{X}) \ ,$$

and $\nabla^2 p_{\mathbf{Y}}(y) = \mathbb{E}\Big[\big(-\frac{1}{\sigma^2} I_d + \frac{1}{\sigma^4}(y-\mathbf{X})(y-\mathbf{X})^\top\big) p_{\mathbf{Y}}(y|\mathbf{X})\Big]$. This gives

$$\begin{aligned}\frac{\nabla^2 p_{\mathbf{Y}}(y)}{p_{\mathbf{Y}}(y)} &= \mathbb{E}\Big[\big(-\frac{1}{\sigma^2} I_d + \frac{1}{\sigma^4}(y-\mathbf{X})(y-\mathbf{X})^\top\big)\frac{p_{\mathbf{Y}}(y|\mathbf{X})}{p_{\mathbf{Y}}(y)}\Big] \ , \\ &= -\frac{1}{\sigma^2} I_d + \frac{1}{\sigma^4}\mathbb{E}[(y-\mathbf{X})(y-\mathbf{X})^\top|\mathbf{Y}=y] \ .\end{aligned}$$

Combined with the result $\frac{\nabla p_{\mathbf{Y}}(y)}{p_{\mathbf{Y}}(y)} = -\frac{1}{\sigma^2}(\mathbb{E}[y-\mathbf{X}|\mathbf{Y}=y])$ from Proposition 3.2.4, we obtain

$$\begin{aligned}\nabla^2 \log p_{\mathbf{Y}}(y) &= \frac{\nabla^2 p_{\mathbf{Y}}(y)}{p_{\mathbf{Y}}(y)} - \frac{\nabla p_{\mathbf{Y}}(y)}{p_{\mathbf{Y}}(y)}\Big(\frac{\nabla p_{\mathbf{Y}}(y)}{p_{\mathbf{Y}}(y)}\Big)^\top \\ &= -\frac{1}{\sigma^2} I_d + \frac{1}{\sigma^4}\mathbb{E}[(y-\mathbf{X})(y-\mathbf{X})^\top|\mathbf{Y}=y] \\ &\quad - \frac{1}{\sigma^4}(\mathbb{E}[y-\mathbf{X}|\mathbf{Y}=y])(\mathbb{E}[y-\mathbf{X}|\mathbf{Y}=y])^\top \\ &= -\frac{1}{\sigma^2} I_d + \frac{1}{\sigma^4}\operatorname{Cov}[y-\mathbf{X}|\mathbf{Y}=y] \ .\end{aligned}$$

With $\operatorname{Cov}[y-\mathbf{X}|\mathbf{Y}=y] = \operatorname{Cov}[\mathbf{X}|\mathbf{Y}=y] = \sigma^2 \operatorname{Cov}[\sigma^{-1}\mathbf{X}|\mathbf{Y}]$, we complete the proof. □

Proof of Proposition 3.2.6. Using the integration by parts, we have

$$\begin{aligned}\int \operatorname{tr}[\nabla^2 \log p_\nu(x)]\mathrm{d}\nu(x) &= \sum_{i=1}^d \int p_\nu(x)\frac{\partial^2}{\partial x_i}\log p_\nu(x)\mathrm{d}x \ , \\ &= -\sum_{i=1}^d \int \frac{\partial}{\partial x_i} p_\nu(x)\frac{\partial}{\partial x_i}\log p_\nu(x)\mathrm{d}x \ , \\ &= -\int \|\nabla \log p_\nu(x)\|^2 p_\nu(x)\mathrm{d}x \ .\end{aligned}$$

□

A.2.2 Proofs for Section 3.3

Proof of Theorem 3.3.1. Take any two $\mu, \nu \in \mathcal{P}_2^r(X)$, we have by convexity of $\mathcal{E}(\cdot)$

$$\mathcal{E}(\mu) - \mathcal{E}(\nu) \geq \int \langle \nabla \log p_\nu, \mathbf{t}_\nu^\mu - \mathbf{i} \rangle \mathrm{d}\nu = \langle \nabla \log p_\nu(\mathbf{t}_\mu^\nu), \mathbf{i} - \mathbf{t}_\mu^\nu \rangle \mathrm{d}\mu \ ,$$
$$\mathcal{E}(\nu) - \mathcal{E}(\mu) \geq \int \langle \nabla \log p_\mu, \mathbf{t}_\mu^\nu - \mathbf{i} \rangle \mathrm{d}\mu \ .$$

Therefore, adding up these two inequalities, we have

$$\int \langle \nabla \log p_\mu - \nabla \log p_\nu(\mathbf{t}_\mu^\nu), \mathbf{t}_\mu^\nu - \mathbf{i} \rangle \mathrm{d}\mu \leq 0 \ . \tag{A.2.2}$$

Now denote $\mathbf{t} := \mathbf{t}_{\mu_t}^{\nu_t}$, we know $(\mathbf{f}^{\mu,\eta}, \mathbf{f}^{\nu,\eta} \circ \mathbf{t})_{\#}\mu_t \in \Pi(\mu_t, \nu_t)$ where $\mathbf{f}^{\mu,\eta}, \mathbf{f}^{\nu,\eta}$ are defined in (3.3.1),

$$\begin{aligned}
W_2^2(\mu_{t+\eta}, \nu_{t+\eta}) &= \inf_{\pi \in \Pi(\mu_{t+\eta}, \nu_{t+\eta})} \int \|x - y\|^2 \mathrm{d}\pi(x, y) \ , \\
&\leq \int \|(\mathbf{i} - \eta \nabla f - \eta \beta^{-1} \nabla \log p_{\mu_t}) - (\mathbf{t} - \eta \nabla f(\mathbf{t}) - \eta \beta^{-1} \nabla \log p_{\nu_t}(\mathbf{t}))\|^2 \mathrm{d}\mu_t \ , \\
&= \int \|\mathbf{i} - \mathbf{t}\|^2 \mathrm{d}\mu_t - 2\eta \int \langle \mathbf{i} - \mathbf{t}, \nabla f - \nabla f(\mathbf{t}) \rangle \mathrm{d}\mu_t \\
&\quad + 2\eta\beta^{-1} \int \langle \nabla \log p_{\mu_t} - \nabla \log p_{\nu_t}(\mathbf{t}), \mathbf{t} - \mathbf{i} \rangle \mathrm{d}\mu_t \\
&\quad + O(\eta^2) \ , \\
&\leq (1 - 2\eta\lambda) W_2^2(\mu_t, \nu_t) + O(\eta^2) \ .
\end{aligned}$$

Here we use (A.2.2) and the fact $\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq \lambda \|x - y\|^2$, for all x, y. Rearrange and then let $\eta \to 0$, we prove the theorem. □

Proof of Theorem 3.3.3. First, for any $\nu \in \mathcal{P}_2^r(X)$, define two quantities

$$\epsilon := \|\mathbf{t}_{\mu_t}^\nu - \mathbf{i}\|_{L^2(\mu_t; \mathbb{R}^d)} = W_2(\nu, \mu_t) \ ,$$
$$\boldsymbol{\xi} := (\mathbf{t}_{\mu_t}^\nu - \mathbf{i}) / \|\mathbf{t}_{\mu_t}^\nu - \mathbf{i}\|_{L^2(\mu_t; \mathbb{R}^d)} \ ,$$

Then $\mathbf{t}_{\mu_t}^{\nu} = \mathbf{i} + \epsilon\boldsymbol{\xi}$. For any $\nu \in \mathcal{P}_2^r(X)$, we have

$$\begin{aligned} W_2^2(\mathbf{b}_{\#}^{\mu,\eta}\mu_t, \mathbf{b}_{\#}^{\mu,\eta}\nu) &\leq \int \|\mathbf{b}^{\mu,\eta} \circ (\mathbf{i} + \epsilon\boldsymbol{\xi}) - \mathbf{b}^{\mu,\eta} \circ \mathbf{i}\|^2 \mathrm{d}\mu_t \ , \\ &= \int \|\mathbf{b}^{\mu,\eta}(x + \epsilon\boldsymbol{\xi}(x)) - \mathbf{b}^{\mu,\eta}(x)\|^2 \mathrm{d}\mu_t(x) \ . \end{aligned}$$

Define an auxiliary function $g_{\boldsymbol{\xi}}(\epsilon) := \int \|\mathbf{b}^{\mu,\eta}(x + \epsilon\boldsymbol{\xi}(x)) - \mathbf{b}^{\mu,\eta}(x)\|^2 \mathrm{d}\mu_t(x)$, we can verify that

$$g'_{\boldsymbol{\xi}}(0) = 0, \ \lim_{\epsilon \to 0} \frac{g_{\boldsymbol{\xi}}(\epsilon)}{\epsilon} = 0 \ .$$

We also know that

$$\begin{aligned} g'_{\boldsymbol{\xi}}(\epsilon) &= \int 2\langle \nabla \mathbf{b}^{\mu,\eta}(x + \epsilon\boldsymbol{\xi}(x))\boldsymbol{\xi}(x), \mathbf{b}^{\mu,\eta}(x + \epsilon\boldsymbol{\xi}(x)) - \mathbf{b}^{\mu,\eta}(x)\rangle \mathrm{d}\mu_t(x) \ , \\ &\leq \frac{\int \|\mathbf{b}^{\mu,\eta}(x + \epsilon\boldsymbol{\xi}(x)) - \mathbf{b}^{\mu,\eta}(x)\|^2 \mathrm{d}\mu_t(x)}{\epsilon} + \epsilon \cdot \int \|\nabla \mathbf{b}^{\mu,\eta}(x + \epsilon\boldsymbol{\xi}(x))\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) \ , \\ &\leq \frac{g_{\boldsymbol{\xi}}(\epsilon)}{\epsilon} + \epsilon \cdot \sup_x \|\nabla \mathbf{b}^{\mu,\eta}(x)\|_{\mathrm{op}}^2 \cdot \int \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) \ . \end{aligned}$$

Notice $\int \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) = 1$, and divide both sides by ϵ, we obtain

$$\frac{\mathrm{d}}{\mathrm{d}\epsilon}\left(\frac{g_{\boldsymbol{\xi}}(\epsilon)}{\epsilon}\right) = \frac{g'_{\boldsymbol{\xi}}(\epsilon)\epsilon - g_{\boldsymbol{\xi}}(\epsilon)}{\epsilon^2} \leq \sup_x \|\nabla \mathbf{b}^{\mu,\eta}(x)\|_{\mathrm{op}}^2 \ .$$

Integrate this inequality and recall $\lim_{\epsilon \to 0} \frac{g_{\boldsymbol{\xi}}(\epsilon)}{\epsilon} = 0$, we get

$$\frac{g_{\boldsymbol{\xi}}(\epsilon)}{\epsilon} \leq \epsilon \cdot \sup_x \|\nabla \mathbf{b}^{\mu,\eta}(x)\|_{\mathrm{op}}^2 \ ,$$

which implies

$$g_{\boldsymbol{\xi}}(\epsilon) \leq \sup_x \|\nabla \mathbf{b}^{\mu,\eta}(x)\|_{\mathrm{op}}^2 \cdot \epsilon^2 \int \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) = \sup_x \|\nabla \mathbf{b}^{\mu,\eta}(x)\|_{\mathrm{op}}^2 \cdot W_2^2(\mu_t, \nu) \ .$$

Recall the definition of $g_{\boldsymbol{\xi}}(\epsilon)$, we have shown for any $\nu \in \mathcal{P}_2^r(X)$,

$$W_2^2(\mathbf{b}_{\#}^{\mu,\eta}\mu_t, \mathbf{b}_{\#}^{\mu,\eta}\nu) \le g_{\boldsymbol{\xi}}(\epsilon) \le \sup_x \|\nabla \mathbf{b}^{\mu,\eta}(x)\|_{\mathrm{op}}^2 \cdot W_2^2(\mu_t, \nu) \ . \tag{A.2.3}$$

Therefore, we have for any $\nu \in \mathcal{P}_2^r(X)$

$$\frac{W_2^2(\mathbf{b}_{\#}^{\mu,\eta}\mu_t, \mathbf{b}_{\#}^{\mu,\eta}\nu) - W_2^2(\mu_t, \nu)}{W_2^2(\mu_t, \nu)} \le \sup_x \|\nabla \mathbf{b}^{\mu,\eta}(x)\|_{\mathrm{op}}^2 - 1 \ .$$

To further bound the right-hand side, for η small enough, we notice

$$\nabla \mathbf{b}^{\mu,\eta}(x) = I_d + \eta(\nabla^2 f(x) + \beta^{-1}\nabla^2 \log p_{\mu_t}(x)) \preceq \big(1 + \eta(\kappa - \beta^{-1}\zeta)\big) I_d \ .$$

We complete the proof by taking the $\eta \to 0$ limit. □

Proof of Corollary 3.3.5. First note

$$\begin{aligned} \nabla \mathbf{b}^{\mu,\eta}(x) = I_d + \eta(\nabla^2 f(x) + \beta^{-1}\nabla^2 \log p_{\mu_t}(x)) = \big(1 + \eta(\kappa - \beta^{-1}\zeta)\big) I_d \ , \\ \mathbf{b}^{\mu,\eta} = (1 + \eta(\kappa - \beta^{-1}\zeta))\mathbf{i}, \quad \mathbf{f} := (\mathbf{b}^{\mu,\eta})^{-1} = (1 + \eta(\kappa - \beta^{-1}\zeta))^{-1}\mathbf{i} \ . \end{aligned}$$

Define for convenience, $\mu_{t-\eta} := \mathbf{b}_{\#}^{\mu,\eta}\mu_t, \nu_{-\eta} := \mathbf{b}_{\#}^{\mu,\eta}\nu$, then in the proof of Theorem 3.3.3, (A.2.3) already proved

$$\frac{W_2^2(\mu_t, \nu)}{W_2^2(\mathbf{b}_{\#}^{\mu,\eta}\mu_t, \mathbf{b}_{\#}^{\mu,\eta}\nu)} = \frac{W_2^2(\mathbf{f}_{\#}\mu_{t-\eta}, \mathbf{f}_{\#}\nu_{-\eta})}{W_2^2(\mu_{t-\eta}, \nu_{-\eta})} \le \sup_x \|\nabla \mathbf{f}(x)\|_{\mathrm{op}}^2 = (1 + \eta(\kappa - \beta^{-1}\zeta))^{-2} \ .$$

Thus taking the reciprocal, then taking the limit inferior as $\eta \to 0$, we obtain

$$\liminf_{\eta \to 0} \frac{1}{\eta} \frac{W_2^2(\mathbf{b}_{\#}^{\mu,\eta}\mu_t, \mathbf{b}_{\#}^{\mu,\eta}\nu) - W_2^2(\mu_t, \nu)}{W_2^2(\mu_t, \nu)} \ge 2(\kappa - \beta^{-1}\zeta) \ .$$

Theorem 3.3.3 already proved

$$\limsup_{\eta\to 0}\frac{1}{\eta}\frac{W_2^2(\mathbf{b}^{\mu,\eta}_{\#}\mu_t,\mathbf{b}^{\mu,\eta}_{\#}\nu)-W_2^2(\mu_t,\nu)}{W_2^2(\mu_t,\nu)}\le 2(\kappa-\beta^{-1}\zeta)\ .$$

Thus the equality holds. □

Proof of Theorem 3.3.6. Note by definition of the diffuse-then-denoise step, we have

$$\begin{aligned}
&W_2^2(\mathbf{b}^{\mu,\eta}_{\#}\mathbf{f}^{\mu,\eta}_{\#}\mu,\mu)\\
&\le\int\|\mathbf{b}^{\mu,\eta}\circ\mathbf{f}^{\mu,\eta}(x)-x\|^2\mathrm{d}\mu\ ,\\
&=\int\|\mathbf{f}^{\mu,\eta}(x)+\eta(\nabla f(\mathbf{f}^{\mu,\eta}(x))+\beta^{-1}\nabla\log p_{\mu_\eta}(\mathbf{f}^{\mu,\eta}(x)))-x\|^2\mathrm{d}\mu\ ,\\
&=\int\|\mathbf{f}^{\mu,\eta}(x)+\eta\nabla f(\mathbf{f}^{\mu,\eta}(x))+\eta\beta^{-1}\nabla\log p_{\mu}(x)+O(\eta^2)-x\|^2\mathrm{d}\mu\ ,\\
&=\int\|x-\eta\nabla f(x)-\eta\beta^{-1}\nabla\log p_{\mu}(x)+\eta\nabla f(x)+\eta\beta^{-1}\nabla\log p_{\mu}(x)+O(\eta^2)-x\|^2\mathrm{d}\mu\ ,\\
&=O(\eta^4)\ ,
\end{aligned}$$

where the second to the third line is applying the change of variable formula

$$\begin{aligned}
\log p_{\mu_\eta}(\mathbf{f}^{\mu,\eta}(x))&=\log p_\mu(x)-\log\det\Big(I_d-\eta(\nabla^2 f(x)+\beta^{-1}\nabla^2\log p_\mu(x))\Big)\ ,\\
&=\log p_\mu(x)+\eta\,\mathrm{tr}(\nabla^2 f(x)+\beta^{-1}\nabla^2\log p_\mu(x))+O(\eta^2)\ .
\end{aligned}$$

and the third to the fourth line uses the fact

$$\nabla f(\mathbf{f}^{\mu,\eta}(x))=\nabla f(x)+\eta\nabla^2 f(x)(\nabla f(x)+\beta^{-1}\nabla\log p_\mu(x))+O(\eta^2)\ .$$

The proof is completed. □

Proof of Corollary 3.3.7. Without loss of generality, assume $W_2(\mu,\nu)=1$.

By Theorem 3.3.1, we know

$$W_2(\mu_{K\eta}, \nu_{K\eta}) \leq 1 - \eta K + O(\eta^2) \leq \exp\Big(-\eta K + O(\eta^2)\Big) \, .$$

Using Theorem 3.3.3

$$\begin{aligned}
\frac{W_2((\mathbf{b}_K^{\mu,\eta})_{\#}\mu_{K\eta}, (\mathbf{b}_K^{\mu,\eta})_{\#}\nu_{K\eta})}{W_2(\mu_{K\eta}, \nu_{K\eta})} &\leq 1 + \eta(1 - \beta^{-1}\zeta_{K\eta}) + O(\eta^2), \\
&\leq \exp\Big(\eta(1 - \beta^{-1}\zeta_{K\eta}) + O(\eta^2)\Big) \, , \\
W_2((\mathbf{b}_K^{\mu,\eta})_{\#}\mu_{K\eta}, (\mathbf{b}_K^{\mu,\eta})_{\#}\nu_{K\eta})) &\leq \exp\Big(-\eta(K-1) - \eta\beta^{-1}\zeta_{K\eta} + O(\eta^2)\Big) \, .
\end{aligned}$$

Now recall Theorem 3.3.6, we have

$$W_2(\mu_{(K-1)\eta}, (\mathbf{b}_K^{\mu,\eta})_{\#}\mu_{K\eta}) = W_2(\mu_{K-1}, (\mathbf{b}_K^{\mu,\eta} \circ \mathbf{f}_{K-1}^{\mu,\eta})_{\#}\mu_{(K-1)\eta}) = O(\eta^2) \, ,$$

and thus

$$\begin{aligned}
W_2\Big(\mu_{(K-1)\eta}, (\mathbf{b}_K^{\mu,\eta})_{\#}\nu_{K\eta}\Big) &\leq W_2\Big(\mu_{(K-1)\eta}, (\mathbf{b}_K^{\mu,\eta})_{\#}\mu_{K\eta}\Big) \\
&\quad + W_2\big((\mathbf{b}_K^{\mu,\eta})_{\#}\mu_{K\eta}, (\mathbf{b}_K^{\mu,\eta})_{\#}\nu_{K\eta}\big) \\
&\leq \exp\Big(-\eta(K-1) - \eta\beta^{-1}\zeta_{K\eta} + O(\eta^2)\Big) \, .
\end{aligned}$$

Chaining this bound recursively, we have

$$W_2\big(\mu, \big(\mathbf{b}_{[K]}^{\mu} \circ \mathbf{f}_{[K]}^{\nu}\big)_{\#}\nu\big) = W_2(\mu, (\mathbf{b}_{[K]}^{\mu})_{\#}\nu_{K\eta}) \leq \exp\left(-\eta\beta^{-1}\sum_{k=1}^{K}\zeta_{k\eta} + O(\eta^2)\right) \, .$$

□

A.2.3 Proofs for Section 3.4

Proof of Proposition 3.4.4. Recall the definition

$$\frac{\mathrm{d}m_\mu(\delta,\mathsf{r})}{\mathrm{d}\delta} = \frac{s_\mathsf{r}(1-\delta)\cdot\delta - \int_{1-\delta}^{\infty} s_\mathsf{r}(u)\mathrm{d}u}{\delta^2} \,.$$

(1) In the log-concave case, we know $s_\mathsf{r}(1)=0$, therefore

$$\frac{\mathrm{d}m_\mu(\delta,\mathsf{r})}{\mathrm{d}\delta} = \frac{s_\mathsf{r}(1-\delta)\cdot\delta - \int_{1-\delta}^{1} s_\mathsf{r}(u)\mathrm{d}u}{\delta^2} = \frac{\int_{1-\delta}^{1}[s_\mathsf{r}(1-\delta)-s_\mathsf{r}(u)]\mathrm{d}u}{\delta^2} \geq 0$$

which shows the non-decreasing shape of $m_\mu(\cdot,\mathsf{r})$. (2) In the non-log-concave case, we have $s_\mathsf{r}(1)>0$, and thus

$$\frac{\mathrm{d}m_\mu(\delta,\mathsf{r})}{\mathrm{d}\delta}|_{\delta\to 0+} < 0$$

and we claim that (by taking the derivative w.r.t δ)

$$\delta \mapsto s_\mathsf{r}(1-\delta)\cdot\delta - \int_{1-\delta}^{\infty} s_\mathsf{r}(u)\mathrm{d}u$$

is non-decreasing in δ. Therefore, $\delta \mapsto \frac{\mathrm{d}m_\mu(\delta,\mathsf{r})}{\mathrm{d}\delta}$ either crosses zero (in a non-decreasing way) or stays negative for $\delta\in[0,1]$. Therefore, $m_\mu(\delta,\mathsf{r})$ is either (i) non-increasing in $\delta\in(0,1]$, or (ii) U-shaped in $\delta\in(0,1]$, namely first non-increasing then non-decreasing. □

Proof of Theorem 3.4.6. As in the proof of Theorem 3.3.3, for any $\nu\in\mathcal{M}(\mu_t,M)$, define two quantities

$$\epsilon := \|\mathbf{t}_{\mu_t}^{\nu} - \mathbf{i}\|_{L^2(\mu_t;\mathbb{R}^d)} = W_2(\nu,\mu_t) \,,$$
$$\boldsymbol{\xi} := (\mathbf{t}_{\mu_t}^{\nu} - \mathbf{i})/\|\mathbf{t}_{\mu_t}^{\nu} - \mathbf{i}\|_{L^2(\mu_t;\mathbb{R}^d)} \,,$$

Then $\mathbf{t}_{\mu_t}^{\nu} = \mathbf{i} + \epsilon\boldsymbol{\xi}$, and as $\nu \overset{W_2}{\to} \mu_t$, we know $\epsilon\to 0$. Now note additionally if we assume

$\nu \in \mathcal{M}(\mu_t, M)$, then the corresponding $\boldsymbol{\xi}$ satisfies

$$\boldsymbol{\xi} \in \mathcal{T}_{\mu_t}(M) := \left\{ \boldsymbol{\xi} \ : \ \int \|\boldsymbol{\xi}\|^2 \mathrm{d}\mu_t = 1, \ \sup_{x \in \mathrm{Dom}(\mu_t)} \|\boldsymbol{\xi}(x)\|^2 \leq M \right\} .$$

Claim that for any sequence of $\nu \in \mathcal{P}_2^r(X) : \nu \overset{W_2}{\to} \mu_t$ that attains the limit superior, we have

$$\begin{aligned} &\limsup_{\nu \in \mathcal{M}(\mu_t, M) : \nu \overset{W_2}{\to} \mu_t} \frac{W_2^2(\mathbf{b}_{\#}^{\mu,\eta}\mu_t, \mathbf{b}_{\#}^{\mu,\eta}\nu) - W_2^2(\mu_t, \nu)}{W_2^2(\mu_t, \nu)} \\ &\leq \sup_{\boldsymbol{\xi} \in \mathcal{T}_{\mu_t}(M)} \frac{\int \|\nabla \mathbf{b}^{\mu,\eta}(x)\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) - \int \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x)}{\int \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x)} . \end{aligned}$$

To derive this claim, notice that

$$\begin{aligned} W_2^2(\mathbf{b}_{\#}^{\mu,\eta}\mu_t, \mathbf{b}_{\#}^{\mu,\eta}\nu) &\leq \int \|\mathbf{b}^{\mu,\eta} \circ (\mathbf{i} + \epsilon\boldsymbol{\xi}) - \mathbf{b}^{\mu,\eta} \circ \mathbf{i}\|^2 \mathrm{d}\mu_t , \\ &= \int \|\mathbf{b}^{\mu,\eta}(x + \epsilon\boldsymbol{\xi}(x)) - \mathbf{b}^{\mu,\eta}(x)\|^2 \mathrm{d}\mu_t(x) , \\ &= \epsilon^2 \left(\int \|\nabla \mathbf{b}^{\mu,\eta}(x)\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) + o_\epsilon(1) \right) . \end{aligned}$$

The last step requires some justification. Define an auxiliary function $g(\epsilon) := \int \|\mathbf{b}^{\mu,\eta}(x + \epsilon\boldsymbol{\xi}(x)) - \mathbf{b}^{\mu,\eta}(x)\|^2 \mathrm{d}\mu_t(x)$, we can verify that

$$\begin{aligned} &g'(0) = 0, \\ &g''(0) = 2 \int \|\nabla \mathbf{b}^{\mu,\eta}(x)\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) , \\ &g''(\epsilon) = 2 \int \|\nabla \mathbf{b}^{\mu,\eta}(x + \epsilon\boldsymbol{\xi}(x))\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) \\ &\qquad + 2 \int \left\langle \nabla^2 \mathbf{b}^{\mu,\eta}(x + \epsilon\boldsymbol{\xi}(x)), \boldsymbol{\xi}(x) \otimes \boldsymbol{\xi}(x) \otimes \left(\mathbf{b}^{\mu,\eta}(x + \epsilon\boldsymbol{\xi}(x)) - \mathbf{b}^{\mu,\eta}(x) \right) \right\rangle . \end{aligned}$$

By the Taylor's Theorem, there exists $\tilde{\epsilon} \in [0, \epsilon]$, such that

$$g(\epsilon) = \frac{1}{2} g''(\tilde{\epsilon}) \epsilon^2 \ .$$

Notice $g''(\tilde{\epsilon}) = g''(0) + o_\epsilon(1)$ due to the fact that $\|\boldsymbol{\xi}(x)\|^2 \leq M$ uniformly bounded and that $\mathbf{b}^{\mu,\eta}$ has bounded derivatives, we establish the claim.

Now we proceed to control the term $\int \|\nabla \mathbf{b}^{\mu,\eta}(x)\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x)$

$$\int \|\nabla \mathbf{b}^{\mu,\eta}(x)\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) \leq \int \|\nabla \mathbf{b}^{\mu,\eta}(x)\|_{\mathrm{op}}^2 \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) \ .$$

Now let's study $\|\nabla \mathbf{b}^{\mu,\eta}(x)\|_{\mathrm{op}}$: recall $\mathsf{s}(t) = \sqrt{\beta^{-1}(1 - e^{-2t})}$

$$\begin{aligned}
\|\nabla \mathbf{b}^{\mu,\eta}(x)\|_{\mathrm{op}} &= \|I_d + \eta I_d - \tfrac{\eta}{1-e^{-2t}}(I_d - \mathrm{Cov}[\mathsf{r}(t)\mathbf{X}_0 | \mathbf{X}_t = x])\|_{\mathrm{op}} \\
&= (1+\eta) - \tfrac{\eta}{1-e^{-2t}} + \tfrac{\eta}{1-e^{-2t}} \|\mathrm{Cov}[\mathsf{r}(t)\mathbf{X}_0 | \mathbf{X}_t = x]\|_{\mathrm{op}} \\
&= (1+\eta) - \tfrac{\eta}{1-e^{-2t}} + \tfrac{\eta}{1-e^{-2t}} L_{\mathsf{r}(t)}(\tfrac{x}{\mathsf{s}(t)})
\end{aligned}$$

Here we recall the conditional covariance (localization function) defined before

$$L_{\mathsf{r}}(y) := \|\mathrm{Cov}[\mathsf{r}\mathbf{X} | \mathbf{Y}_{\mathsf{r}} = y]\|_{op} \ .$$

Continue, we have

$$\begin{aligned}
&\int \|\nabla \mathbf{b}^{\mu,\eta}(x)\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) - \int \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) \\
&\leq 2\eta \int \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) - \tfrac{2\eta}{1-e^{-2t}} \int \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) \\
&\quad + \tfrac{2\eta}{1-e^{-2t}} \int L_{\mathsf{r}(t)}(\tfrac{x}{\mathsf{s}(t)}) \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) + O(\eta^2) \ .
\end{aligned}$$

We analyze the last term: first, we define the region

$$R_\delta := \{x \in X \ : \ L_{\mathsf{r}(t)}(\tfrac{x}{\mathsf{s}(t)}) \le 1-\delta\}$$

and bound the integral depending on the region,

$$\begin{aligned}
\int L_{\mathsf{r}(t)}(\tfrac{x}{\mathsf{s}(t)})\|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) &= \int_{x\in R_\delta} L_{\mathsf{r}(t)}(\tfrac{x}{\mathsf{s}(t)})\|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) \\
&\quad + \int_{x\in R_\delta^c} L_{\mathsf{r}(t)}(\tfrac{x}{\mathsf{s}(t)})\|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) \ , \\
&\le (1-\delta)\left(\int_{x\in R_\delta} \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) + \int_{x\in R_\delta^c} \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x)\right) \\
&\quad + \int_{x\in R_\delta^c} \big(L_{\mathsf{r}(t)}(\tfrac{x}{\mathsf{s}(t)}) - (1-\delta)\big)\|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) \ ,
\end{aligned}$$

Recalling $\boldsymbol{\xi} \in \mathcal{T}_{\mu_t}(M)$,

$$\begin{aligned}
&(1-\delta)\int \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) + \int_{x\in R_\delta^c} \big(L_{\mathsf{r}(t)}(\tfrac{x}{\mathsf{s}(t)}) - (1-\delta)\big)\|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) \\
&= (1-\delta)\int \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) + \int \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) \cdot M \cdot \int_{x\in R_\delta^c} \big(L_{\mathsf{r}(t)}(\tfrac{x}{\mathsf{s}(t)}) - (1-\delta)\big)\mathrm{d}\mu_t(x) \ , \\
&= \int \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) \cdot \left((1-\delta) + M \cdot \int_{1-\delta}^\infty s_{\mathsf{r}(t)}(z)\mathrm{d}z\right) \ ,
\end{aligned}$$

where the last step uses Proposition 3.2.6 and Fubini's theorem: for a non-negative random variable $Z > 0$, if $\mathbb{E}[Z] < \infty$, then $\int_C^\infty (z-C)p_Z(z)\mathrm{d}z = \int_C^\infty P(Z \ge z)\mathrm{d}z$ for $C > 0$; set $Z = L_{\mathsf{r}(t)}(\frac{\mathbf{X}_t}{\mathsf{s}(t)})$ and $C = 1-\delta$. Put things together, for any $\delta \in [0,1]$

$$\begin{aligned}
&\sup_{\boldsymbol{\xi}\in\mathcal{T}_{\mu_t}(M)} \frac{\int \|\nabla \mathbf{b}^{\mu,\eta}(x)\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x) - \int \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x)}{\int \|\boldsymbol{\xi}(x)\|^2 \mathrm{d}\mu_t(x)} \\
&\le 2\eta - \tfrac{2\eta}{1-e^{-2t}}\big[\delta - M \cdot h_\mu(\delta, \mathsf{r}(t))\big] + O(\eta^2) \ .
\end{aligned}$$

□

APPENDIX B

APPENDICES TO CHAPTER 5

B.1 Related literature

This work sits at the intersection of four literatures: conditional moment restrictions, nonparametric instrumental variables, machine-learning approaches to IV, and simulation-based econometrics.

Our framework belongs to the general literature on estimation under conditional moment restrictions. Chamberlain [1987] gives the foundational efficiency perspective, and Ai and Chen [2003] develop efficient semiparametric estimation for conditional moment models with unknown functions. Instrumental variables is a leading special case. In the nonlinear setting, the identifying restriction induces a nonparametric inverse problem. This point is emphasized by Newey and Powell [2003], Hall and Horowitz [2005], and Darolles et al. [2011], who develop nonparametric IV estimators based on sieve, kernel, and regularization methods. Our work shares the same identifying restriction, but takes a different computational route. Rather than regularizing the inverse problem directly, we learn a conditional first-stage law and use it to evaluate the conditional moment restriction by simulation.

Our work is related to the literature on nonparametric IV estimation under ill-posedness. Existing methods typically proceed by regularizing an operator inversion or by approximating the conditional restriction with a growing collection of unconditional moments. These approaches are well understood and theoretically well founded, but their performance depends on basis choice, bandwidth or regularization tuning, and the severity of the inverse problem; see Newey and Powell [2003], Hall and Horowitz [2005], Darolles et al. [2011]. This becomes especially salient when the first-stage law is high-dimensional, strongly nonlinear, or conditionally multimodal. Our procedure is designed for precisely those environments. The first stage is treated as a conditional generative problem, and the second stage works directly with an empirical analogue of the conditional moment discrepancy.

There has been a collection of recent works that have explored the plausibility of machine-learning approaches to tackle the nonlinear IV problem. The closest connection is to DeepIV [Hartford et al., 2017], which also learns a flexible first-stage distribution and uses Monte Carlo integration in the second stage. The difference is in the target criterion. DeepIV estimates the structural function through a prediction-risk objective averaged over the learned first-stage law, whereas our estimator minimizes an empirical analogue of the IV conditional moment discrepancy itself. In that sense, our procedure remains closer to the identifying restriction. More broadly, our work is related to kernel and adversarial approaches to conditional moment problems, including kernel IV regression [Singh et al., 2019] and minimax or adversarial estimators such as Dikkala et al. [2020]. Those methods enlarge the function classes used on either the structural side or the moment side. Our contribution is complementary: we enlarge the class of admissible first-stage laws by using a flexible conditional sampler and then exploit that sampler as a Monte Carlo device inside the structural estimation problem.

This work is also is related to a growing body of simulation-based methods econometrics. In methods of simulated moments and indirect inference, simulation is used to evaluate moments or auxiliary criteria that are analytically inconvenient or intractable; see Pakes and Pollard [1989], McFadden [1989], Gourieroux et al. [1993]. Our procedure follows the same broad logic. The difference is that the simulator here is learned from data rather than imposed by a structural model. In this respect, the procedure may be viewed as a conditional generative analogue of classical simulation-based estimation.

Finally, this chapter is also connected to the broader idea that flexible learned samplers may serve as statistical objects rather than merely as sampling devices. The preceding chapter develops that viewpoint for diffusion models, and the present chapter instantiates it in an IV setting. The learned sampler is not itself the estimand. It is the object that renders the conditional first stage tractable enough for Monte Carlo evaluation of the identifying restriction. In a similar spirit, recent work has looked at optimal couplings and measure transport

methods as playing an instrumental role for counterfactual imputation and individualized inference problems [Hur and Liang, 2024].

B.2 Classical Approaches to Nonlinear IV

This appendix reviews several standard approaches to IV estimation through the lens of the conditional moment restriction

$$\mathbb{E}[Y - g_0(X, W) \mid Z, W] = 0.$$

The purpose is not to provide a comprehensive survey, but to clarify the sense in which existing estimators make the conditional restriction tractable. Linear IV and sieve IV replace the conditional restriction by finitely many unconditional moments. NPIV methods work more directly with the associated conditional expectation operator, but require regularization of an ill-posed inverse problem. Control-function methods impose additional structure on the first stage. The generative IV estimator developed in the main text takes a different route: it uses a learned conditional sampler to approximate the conditional expectations entering the moment restriction.

Linear IV and two-stage least squares. Consider the linear structural specification

$$Y = X^\top \beta_0 + W^\top \gamma_0 + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid Z, W] = 0. \tag{B.2.1}$$

Let

$$R := (X^\top, W^\top)^\top, \qquad Q := (Z^\top, W^\top)^\top, \qquad \theta_0 := (\beta_0^\top, \gamma_0^\top)^\top.$$

Then the conditional moment restriction implies the unconditional moment condition

$$\mathbb{E}\Big[Q(Y - R^\top \theta_0)\Big] = 0. \tag{B.2.2}$$

When the model is overidentified, the population 2SLS coefficient is

$$\theta_0 = \left\{\mathbb{E}[RQ^\top]\mathbb{E}[QQ^\top]^{-1}\mathbb{E}[QR^\top]\right\}^{-1}\mathbb{E}[RQ^\top]\mathbb{E}[QQ^\top]^{-1}\mathbb{E}[QY], \tag{B.2.3}$$

under the usual rank conditions. In the exactly identified case, this reduces to the simpler inverse of the moment matrix.

In sample, letting R and Q also denote the $N \times d_R$ and $N \times d_Q$ design matrices, the 2SLS estimator is

$$\widehat{\theta}_{\text{2SLS}} = (R^\top P_Q R)^{-1} R^\top P_Q Y, \qquad P_Q := Q(Q^\top Q)^{-1}Q^\top. \tag{B.2.4}$$

Equivalently, 2SLS projects the regressors R onto the span of the instruments and exogenous controls Q, and then regresses Y on that projected component.

From the perspective of the conditional moment restriction, 2SLS performs two reductions. First, it restricts the structural function to be linear in R. Second, it imposes the conditional restriction only through the finite-dimensional instrument vector Q. This gives a stable and interpretable estimator, but the price is a globally linear structural specification and a finite-dimensional representation of the identifying information in (Z, W).

Sieve 2SLS. Sieve 2SLS extends this projection logic to nonlinear structural functions. Let $p_K(X, W)$ be a K-dimensional basis for the structural function and $q_L(Z, W)$ an L-dimensional basis for the instrument space. The structural function is approximated by

$$g_0(X, W) \approx p_K(X, W)^\top \theta, \tag{B.2.5}$$

and θ is estimated from the unconditional moments

$$\mathbb{E}\Big[q_L(Z, W)\{Y - p_K(X, W)^\top \theta\}\Big] = 0. \tag{B.2.6}$$

If P and Q denote the sample matrices with rows $p_K(X_i, W_i)^\top$ and $q_L(Z_i, W_i)^\top$, the sample estimator is

$$\widehat{\theta}_{K,L} = \left\{ P^\top Q (Q^\top Q)^{-1} Q^\top P \right\}^{-1} P^\top Q (Q^\top Q)^{-1} Q^\top Y. \tag{B.2.7}$$

The fitted structural function is

$$\widehat{g}_{K,L}(x, w) = p_K(x, w)^\top \widehat{\theta}_{K,L}.$$

Sieve IV can approximate nonlinear structural functions while retaining a least-squares form. Its performance, however, depends on both the structural basis p_K and the instrument basis q_L. Even when K and L grow with the sample size, estimation proceeds through a finite-dimensional approximation to the conditional moment restriction. Thus the estimator depends on how well the chosen instrument functions represent the identifying variation in the conditional law of X given (Z, W); see Newey and Powell [2003], Blundell et al. [2007], and Darolles et al. [2011].

The inverse-problem view. The conditional moment restriction can also be written as an operator equation. Define

$$(Tg)(z, w) := \mathbb{E}[g(X, W) \mid Z = z, W = w], \qquad m_0(z, w) := \mathbb{E}[Y \mid Z = z, W = w].$$

Then g_0 solves

$$Tg_0 = m_0. \tag{B.2.8}$$

In nonlinear IV, T is typically a smoothing operator, so recovering g_0 from m_0 is an ill-posed inverse problem.

The scalar case makes this transparent. Suppose there are no covariates and

$$Y = g_0(X) + \varepsilon, \qquad \mathbb{E}[\varepsilon \mid Z] = 0.$$

If $X \mid Z = z$ has density $f_{X|Z}(\cdot \mid z)$, then

$$m_0(z) = \mathbb{E}[Y \mid Z = z] = \int g_0(x) f_{X|Z}(x \mid z)\, dx. \tag{B.2.9}$$

This is a Fredholm integral equation of the first kind. For example, if

$$X = Z + V, \qquad V \sim N(0, \sigma^2), \qquad V \perp Z,$$

then

$$m_0(z) = \int g_0(x)\phi_\sigma(x - z)\, dx = (g_0 * \phi_\sigma)(z), \tag{B.2.10}$$

where ϕ_σ is the $N(0, \sigma^2)$ density. Thus T convolves g_0 with a Gaussian kernel. Recovering g_0 requires deconvolution, a canonical ill-posed problem.

Regularized NPIV methods address this instability by replacing the inverse problem with a penalized or truncated approximation. A representative Tikhonov formulation is

$$\widehat{g}_\lambda \in \arg\min_{g \in \mathcal{G}} \|\widehat{T} g - \widehat{m}\|^2 + \lambda \|g\|^2, \tag{B.2.11}$$

where $\widehat{T}$ and $\widehat{m}$ are estimated first-stage objects and λ controls regularization. Other NPIV estimators regularize through sieve truncation or kernel methods. These procedures are general, but their finite-sample behavior depends on the severity of ill-posedness and the choice of regularization; see Hall and Horowitz [2005], Darolles et al. [2011], and Horowitz [2011].

Relation to generative IV. The approaches above make the IV problem tractable in different ways. Linear IV and sieve IV project the conditional moment restriction onto a finite-dimensional instrument space. Regularized NPIV works with the conditional expectation operator but stabilizes its inversion. Control-function methods impose triangular structure that converts endogeneity into a conditioning problem. Generative IV instead

learns a conditional law for the reduced-form variables and uses it as a Monte Carlo device for evaluating the conditional moment criterion. This does not remove the need for identification or first-stage accuracy. Rather, it changes the computational representation of the conditional moment restriction: the key first-stage object is the ability to integrate the residual class accurately under $P(\cdot \mid Z, W)$.

B.3 Technical Proofs

B.3.1 Proofs for Section 5.3

Monte Carlo Identities for the Corrected Criterion

This subsection collects the Monte Carlo identities used in Section 5.3.2. The identities formalize the diagonal terms introduced by the plugin square and show how independent conditional batches remove them.

Lemma B.3.1 (Monte Carlo identities for a support-point criterion)**.** *Fix support points $s_i = (z_i, w_i)$, $i = 1, \dots, n$, and let $Q(\cdot \mid S = s)$ be any conditional law for $(Y, X) \mid S = s$. For each i, let*

$$(Y_{ij}, X_{ij}), \qquad j = 1, \dots, m,$$

be i.i.d. draws from $Q(\cdot \mid S = s_i)$, and define

$$\psi_{ij}(\theta) := g_\theta(X_{ij}, w_i) - Y_{ij}, \qquad \bar{\psi}_i(\theta) := \frac{1}{m} \sum_{j=1}^{m} \psi_{ij}(\theta),$$

$$m_i^Q(\theta) := \mathbb{E}_Q[g_\theta(X, w_i) - Y \mid S = s_i], \qquad L_n^Q(\theta) := \frac{1}{n} \sum_{i=1}^{n} m_i^Q(\theta)^2.$$

Let

$$\widehat{L}_{\text{plug}}^Q(\theta) := \frac{1}{n} \sum_{i=1}^{n} \bar{\psi}_i(\theta)^2.$$

If the relevant second moments exist, then

$$\mathbb{E}_Q[\widehat{L}^Q_{\text{plug}}(\theta) \mid s_1, \ldots, s_n, Q] = L^Q_n(\theta) + \frac{1}{nm}\sum_{i=1}^{n} \text{Var}_Q(g_\theta(X, w_i) - Y \mid S = s_i). \tag{B.3.1}$$

Suppose, in addition, that g_θ is differentiable in θ, the relevant moments exist, and differentiation may be interchanged with conditional expectation. Define

$$\dot{\psi}_{ij}(\theta) := \nabla_\theta g_\theta(X_{ij}, w_i), \qquad \bar{\dot{\psi}}_i(\theta) := \frac{1}{m}\sum_{j=1}^{m} \dot{\psi}_{ij}(\theta),$$

$$\dot{m}^Q_i(\theta) := \mathbb{E}_Q[\nabla_\theta g_\theta(X, w_i) \mid S = s_i].$$

Then

$$\nabla_\theta L^Q_n(\theta) = \frac{2}{n}\sum_{i=1}^{n} m^Q_i(\theta)\dot{m}^Q_i(\theta), \tag{B.3.2}$$

and

$$\begin{aligned}\mathbb{E}_Q\Big[\nabla_\theta \widehat{L}^Q_{\text{plug}}(\theta) \mid s_1, \ldots, s_n, Q\Big] &= \nabla_\theta L^Q_n(\theta) \\ &\quad + \frac{2}{nm}\sum_{i=1}^{n} \text{Cov}_Q(g_\theta(X, w_i) - Y,\, \nabla_\theta g_\theta(X, w_i) \mid S = s_i)\,.\end{aligned} \tag{B.3.3}$$

The covariance between a scalar and a vector is understood componentwise.

Finally, let

$$\{(Y_{ij}, X_{ij})\}_{j=1}^m, \qquad \{(Y'_{ij}, X'_{ij})\}_{j=1}^m$$

be independent conditional batches from $Q(\cdot \mid S = s_i)$. Let primed quantities denote averages over the second batch, and define

$$\widehat{L}^Q_U(\theta) := \frac{1}{n}\sum_{i=1}^{n} \bar{\psi}_i(\theta)\bar{\psi}'_i(\theta), \qquad \widehat{G}^Q_U(\theta) := \frac{1}{n}\sum_{i=1}^{n}\{\bar{\dot{\psi}}_i(\theta)\bar{\psi}'_i(\theta) + \bar{\psi}_i(\theta)\bar{\dot{\psi}}'_i(\theta)\}. \tag{B.3.4}$$

Then

$$\nabla_\theta \widehat{L}_U^Q(\theta) = \widehat{G}_U^Q(\theta), \tag{B.3.5}$$

and

$$\mathbb{E}_Q[\widehat{L}_U^Q(\theta) \mid s_1, \ldots, s_n, Q] = L_n^Q(\theta), \tag{B.3.6}$$

$$\mathbb{E}_Q[\widehat{G}_U^Q(\theta) \mid s_1, \ldots, s_n, Q] = \nabla_\theta L_n^Q(\theta). \tag{B.3.7}$$

Proof. Condition on $s_1, \ldots, s_n$ and Q. For a fixed i, write

$$\psi := g_\theta(X, w_i) - Y, \quad \dot{\psi} := \nabla_\theta g_\theta(X, w_i), \quad m_i := \mathbb{E}_Q[\psi \mid S = s_i], \quad \dot{m}_i := \mathbb{E}_Q[\dot{\psi} \mid S = s_i].$$

For the one-batch averages,

$$\mathbb{E}_Q[\bar{\psi}_i^2 \mid S = s_i] = m_i^2 + \frac{1}{m} \operatorname{Var}_Q(\psi \mid S = s_i),$$

and

$$\mathbb{E}_Q[\bar{\psi}_i \bar{\dot{\psi}}_i \mid S = s_i] = m_i \dot{m}_i + \frac{1}{m} \operatorname{Cov}_Q(\psi, \dot{\psi} \mid S = s_i).$$

Averaging these two identities over i gives (B.3.1) and (B.3.3), while differentiating

$$L_n^Q(\theta) = n^{-1} \sum_{i=1}^{n} m_i^Q(\theta)^2$$

gives (B.3.2).

For the independent-batch criterion, direct differentiation gives

$$\nabla_\theta \widehat{L}_U^Q(\theta) = \widehat{G}_U^Q(\theta).$$

Independence of the two conditional batches gives

$$\mathbb{E}_Q[\bar{\psi}_i\bar{\psi}_i' \mid S = s_i] = m_i^2,$$

and

$$\mathbb{E}_Q[\bar{\dot{\psi}}_i\bar{\psi}_i' + \bar{\psi}_i\bar{\dot{\psi}}_i' \mid S = s_i] = 2m_i\dot{m}_i.$$

Averaging over i gives (B.3.6) and (B.3.7). □

Relating the Generator-Induced and Oracle Criteria

This subsection proves Proposition 5.3.1. Recall that

$$L_n^P(\theta) = \frac{1}{n}\sum_{i=1}^n m_\theta^P(s_i)^2, \qquad L_n^{\widehat{P}}(\theta) = \frac{1}{n}\sum_{i=1}^n m_\theta^{\widehat{P}}(s_i)^2.$$

Define

$$\delta_i(\theta) := m_\theta^{\widehat{P}}(s_i) - m_\theta^P(s_i), \qquad D_n(\theta)^2 := \frac{1}{n}\sum_{i=1}^n \delta_i(\theta)^2.$$

Proof of Proposition 5.3.1. Since

$$m_\theta^{\widehat{P}}(s_i) = m_\theta^P(s_i) + \delta_i(\theta),$$

we have

$$L_n^{\widehat{P}}(\theta) = \frac{1}{n}\sum_{i=1}^n \left\{m_\theta^P(s_i) + \delta_i(\theta)\right\}^2.$$

Expanding the square gives the exact identity

$$L_n^{\widehat{P}}(\theta) - L_n^P(\theta) = \frac{2}{n}\sum_{i=1}^n m_\theta^P(s_i)\delta_i(\theta) + D_n(\theta)^2. \tag{B.3.8}$$

By Cauchy–Schwarz,

$$\left|\frac{2}{n}\sum_{i=1}^{n} m_\theta^P(s_i)\delta_i(\theta)\right| \le 2\left(\frac{1}{n}\sum_{i=1}^{n} m_\theta^P(s_i)^2\right)^{1/2}\left(\frac{1}{n}\sum_{i=1}^{n}\delta_i(\theta)^2\right)^{1/2}.$$

Therefore

$$\left|L_n^{\widehat{P}}(\theta) - L_n^P(\theta)\right| \le 2\sqrt{L_n^P(\theta)}\, D_n(\theta) + D_n(\theta)^2. \tag{B.3.9}$$

□

B.3.2 Proofs for Section 5.4

This appendix proves Proposition 5.4.3 and Theorem 5.4.4. Throughout, for a support point $s_i = (z_i, w_i)$, write

$$P_i := P(\cdot \mid S = s_i), \qquad \widehat{P}_i := \widehat{P}(\cdot \mid S = s_i),$$

where both laws are distributions over (Y, X). Recall that

$$g_\theta(x, w) = p_K(x, w)^\top \theta, \qquad \Theta_C := \{\theta \in \mathbb{R}^K : \|\theta\|_2 \le C\},$$

and

$$\psi_\theta(y, x, w) := g_\theta(x, w) - y.$$

Proof of Proposition 5.4.3. Fix $\theta \in \Theta_C$ and $w \in \mathcal{W}_0$. For any (y, x) and (y', x'),

$$\begin{aligned}|\psi_\theta(y, x, w) - \psi_\theta(y', x', w)| &= \left|p_K(x, w)^\top\theta - y - p_K(x', w)^\top\theta + y'\right| \\ &\le |y - y'| + \left|\{p_K(x, w) - p_K(x', w)\}^\top\theta\right|.\end{aligned}$$

By Cauchy–Schwarz and (5.4.3),

$$\left|\{p_K(x, w) - p_K(x', w)\}^\top\theta\right| \le \|\theta\|_2\|p_K(x, w) - p_K(x', w)\|_2 \le CL_p\|x - x'\|.$$

Hence

$$|\psi_\theta(y,x,w) - \psi_\theta(y',x',w)| \le |y-y'| + CL_p\|x-x'\|.$$

Consider the combined vector (y,x),

$$|y-y'| + CL_p\|x-x'\| \le \sqrt{1+C^2L_p^2}\,\|(y,x)-(y',x')\|.$$

Thus each $\psi_\theta(\cdot,\cdot,w)$ is L_C-Lipschitz in (y,x), uniformly over $\theta \in \Theta_C$ and $w \in \mathcal{W}_0$, where

$$L_C := \sqrt{1+C^2L_p^2}.$$

By Kantorovich–Rubinstein duality, for each i and each $\theta \in \Theta_C$,

$$\left|\int \psi_\theta(y,x,w_i)\, d(\widehat{P}_i - P_i)(y,x)\right| \le L_C W_1(\widehat{P}_i, P_i).$$

Squaring, averaging over i, and taking the supremum over $\theta \in \Theta_C$ gives

$$\Delta_{\Psi_C,n}(\widehat{P},P) \le L_C \left\{\frac{1}{n}\sum_{i=1}^{n} W_1(\widehat{P}_i,P_i)^2\right\}^{1/2}.$$

This proves (5.4.5). If the conditional laws have finite second moments, then

$$W_1(\widehat{P}_i,P_i) \le W_2(\widehat{P}_i,P_i),$$

which gives (5.4.6). □

Lemma B.3.2 (Uniform convergence of the oracle support-point criterion)**.** *Suppose $s_1,\dots,s_n$ are i.i.d. draws from the marginal law of S, Θ_C is compact,* (5.4.2) *holds, and*

$$\mathbb{E}[\alpha(S)^2] < \infty, \qquad \alpha(s) := \mathbb{E}[Y \mid S = s].$$

Then

$$\sup_{\theta \in \Theta_C} \left| L_n^P(\theta) - L(\theta) \right| = o_p(1).$$

Proof. For $\theta \in \Theta_C$, write

$$m_\theta(s) = \mathbb{E}[g_\theta(X, W) - Y \mid S = s].$$

Since $S = s = (z, w)$ fixes $W = w$,

$$m_\theta(s) = \mathbb{E}[p_K(X, w)^\top \theta \mid S = s] - \alpha(s).$$

By (5.4.2) and $\|\theta\|_2 \leq C$,

$$|m_\theta(s)| \leq C M_p + |\alpha(s)|.$$

Therefore

$$m_\theta(S)^2 \leq 2C^2 M_p^2 + 2\alpha(S)^2,$$

which is integrable by assumption.

Moreover, for $\theta, \widetilde{\theta} \in \Theta_C$,

$$\begin{aligned} |m_\theta(s) - m_{\widetilde{\theta}}(s)| &= \left| \mathbb{E}[p_K(X, w)^\top (\theta - \widetilde{\theta}) \mid S = s] \right| \\ &\leq M_p \|\theta - \widetilde{\theta}\|_2. \end{aligned}$$

Hence

$$\begin{aligned} |m_\theta(s)^2 - m_{\widetilde{\theta}}(s)^2| &\leq \left(|m_\theta(s)| + |m_{\widetilde{\theta}}(s)| \right) |m_\theta(s) - m_{\widetilde{\theta}}(s)| \\ &\leq 2M_p \{ C M_p + |\alpha(s)| \} \, \|\theta - \widetilde{\theta}\|_2. \end{aligned}$$

Define the integrable envelope

$$A(s) := 2M_p\{CM_p + |\alpha(s)|\}.$$

Fix $\eta > 0$. Since Θ_C is compact, there exists a finite η-net

$$\{\theta_1, \ldots, \theta_J\} \subseteq \Theta_C.$$

For every $\theta \in \Theta_C$, choose $j(\theta)$ such that

$$\|\theta - \theta_{j(\theta)}\|_2 \leq \eta.$$

Then

$$\begin{aligned}\left|L_n^P(\theta) - L(\theta)\right| \leq \left|L_n^P(\theta_{j(\theta)}) - L(\theta_{j(\theta)})\right| \\ + \frac{1}{n}\sum_{i=1}^{n} A(s_i)\eta + \mathbb{E}[A(S)]\eta.\end{aligned}$$

Taking suprema over $\theta \in \Theta_C$ yields

$$\sup_{\theta\in\Theta_C}\left|L_n^P(\theta) - L(\theta)\right| \leq \max_{1\leq j\leq J}\left|L_n^P(\theta_j) - L(\theta_j)\right| + \eta\left\{\frac{1}{n}\sum_{i=1}^{n} A(s_i) + \mathbb{E}[A(S)]\right\}.$$

For each fixed j, the ordinary law of large numbers gives

$$L_n^P(\theta_j) - L(\theta_j) = o_p(1),$$

and because the net is finite,

$$\max_{1\leq j\leq J}\left|L_n^P(\theta_j) - L(\theta_j)\right| = o_p(1).$$

Also, by the law of large numbers,

$$\frac{1}{n}\sum_{i=1}^{n} A(s_i) \to_p \mathbb{E}[A(S)].$$

Thus, for every fixed $\eta > 0$,

$$\limsup_{n\to\infty} \sup_{\theta\in\Theta_C} \left| L_n^P(\theta) - L(\theta) \right| \leq_p 2\eta\mathbb{E}[A(S)].$$

Letting $\eta \downarrow 0$ proves the result. □

Proof of Theorem 1.

Proof of Theorem 5.4.4. Let $\mathbb{P}_n$ denote the joint law of the data used to train $\widehat{P}$, and the support points $s_1, \ldots, s_n$. All $o_p(1)$ and $O_p(1)$ statements in this section are with respect to $\mathbb{P}_n$. By Lemma B.3.2,

$$\sup_{\theta\in\Theta_C} |L_n^P(\theta) - L(\theta)| = o_p(1).$$

By Proposition 5.3.1, for every $\theta \in \Theta_C$,

$$|L_n^{\widehat{P}}(\theta) - L_n^P(\theta)| \leq 2\sqrt{L_n^P(\theta)}\, D_n(\theta) + D_n(\theta)^2,$$

where

$$D_n(\theta)^2 = \frac{1}{n}\sum_{i=1}^{n} \left[m_\theta^{\widehat{P}}(s_i) - m_\theta^P(s_i) \right]^2.$$

By definition of $\Delta_{\Psi_C,n}(\widehat{P}, P)$,

$$\sup_{\theta\in\Theta_C} D_n(\theta) \leq \Delta_{\Psi_C,n}(\widehat{P}, P).$$

Proposition 5.4.3 and the first-stage condition (5.4.7) imply

$$\Delta_{\Psi_C,n}(\widehat{P}, P) = o_p(1).$$

It remains to note that

$$\sup_{\theta\in\Theta_C} L_n^P(\theta) = O_p(1).$$

Indeed,

$$\sup_{\theta\in\Theta_C} L_n^P(\theta) \le \frac{1}{n}\sum_{i=1}^{n}\{CM_p + |\alpha(s_i)|\}^2 = O_p(1),$$

by $\mathbb{E}[\alpha(S)^2] < \infty$ and the law of large numbers. Therefore,

$$\sup_{\theta\in\Theta_C} |L_n^{\widehat{P}}(\theta) - L_n^P(\theta)| = o_p(1).$$

Combining this with Lemma B.3.2 gives

$$\varepsilon_n := \sup_{\theta\in\Theta_C} |L_n^{\widehat{P}}(\theta) - L(\theta)| = o_p(1).$$

Let

$$\widehat{\theta} \in \arg\min_{\theta\in\Theta_C} L_n^{\widehat{P}}(\theta)$$

be a measurable minimizer. Let $\theta^* \in \Theta_0$. Since $\widehat{\theta}$ minimizes $L_n^{\widehat{P}}$ over Θ_C,

$$L_n^{\widehat{P}}(\widehat{\theta}) \le L_n^{\widehat{P}}(\theta^*).$$

Using uniform convergence,

$$L(\widehat{\theta}) \le L_n^{\widehat{P}}(\widehat{\theta}) + \varepsilon_n \le L_n^{\widehat{P}}(\theta^*) + \varepsilon_n \le L(\theta^*) + 2\varepsilon_n = L_0 + 2\varepsilon_n.$$

Since $L(\widehat{\theta}) \geq L_0$, it follows that

$$L(\widehat{\theta}) - L_0 = o_p(1).$$

Now fix $\varepsilon > 0$. By the identification condition in Theorem 5.4.4, there exists

$$\eta(\varepsilon) > 0$$

such that

$$\operatorname{dist}(\theta, \Theta_0) \geq \varepsilon \quad \Longrightarrow \quad L(\theta) - L_0 \geq \eta(\varepsilon).$$

Therefore

$$\mathbb{P}\{\operatorname{dist}(\widehat{\theta}, \Theta_0) \geq \varepsilon\} \leq \mathbb{P}\{L(\widehat{\theta}) - L_0 \geq \eta(\varepsilon)\} \to 0.$$

Thus

$$\operatorname{dist}(\widehat{\theta}, \Theta_0) \to_p 0.$$

If $\Theta_0 = \{\theta_0\}$, then

$$\operatorname{dist}(\widehat{\theta}, \Theta_0) = \|\widehat{\theta} - \theta_0\|_2,$$

and hence

$$\widehat{\theta} \to_p \theta_0.$$

□

B.4 Implementation Details

This appendix summarizes the implementation used in the synthetic experiments and the schooling application. The code follows the two-stage procedure described in Section 5.3: first fit a conditional generator for the reduced-form law, then use conditional Monte Carlo draws from that generator to estimate the structural function by the corrected CMR criterion.

Table B.1: First-stage conditional diffusion hyperparameters.

Setting	Conditional law	N	T	Steps	Epochs
Design 1	$(Y, X) \mid Z$	5000	4.0	64	3000
Design 2	$(Y, X_1, X_2) \mid (Z_1, Z_2)$	5000	4.0	64	3500
Design 3	$(Y, X_1, X_2) \mid (Z_1, Z_2, W_1, W_2)$	5000	4.5	64	4000
Schooling	$(\log wage, education) \mid (nearcollege, W)$	3010	2.0	64	1000

Conditional generative first stage

The first stage is a conditional score-based generative model implemented through the project `DAG_Learner` class. A model name of the form

`generated variables` | `conditioning variables`

specifies the conditional law to be learned. In the synthetic designs, the fitted laws are

$$(Y, X) \mid Z, \qquad (Y, X_1, X_2) \mid (Z_1, Z_2), \qquad (Y, X_1, X_2) \mid (Z_1, Z_2, W_1, W_2),$$

for Designs 1–3, respectively. In the schooling application, the fitted law is

$$(\log wage, education) \mid (nearcollege, W).$$

The diffusion network uses hidden size 64 and 3 hidden layers throughout. Conditional samples are generated using the reverse-time SDE sampler. The main first-stage hyperparameters are reported in Table B.1.

Conditional moments and second-stage estimation

Support points are sampled from the empirical distribution of the conditioning variables. The synthetic experiments use 500 support points in Designs 1–3. The schooling application uses 1000 support points. At each support point, the fitted generator produces $m = 1000$

Table B.2: Second-stage CMR hyperparameters.

Setting	Input dimension	Structural class	Degree	Epochs	Learning rate
Design 1	1	Polynomial network	5	3500	10^{-2}
Design 2	2	Polynomial network	2	3500	10^{-4}
Design 3	4	Polynomial network	2	4500	8×10^{-4}
Schooling	6	Polynomial network	5	1000	10^{-2}

conditional Monte Carlo draws.

For a support point $s_i = (z_i, w_i)$, generated draws are used to estimate the reduced-form target

$$\widehat{\alpha}_i = \frac{1}{m} \sum_{j=1}^{m} Y_{ij},$$

and to evaluate the structural function $g_\theta(X_{ij}, w_i)$. When W enters the structural function, the support-point value w_i is repeated across the generated treatment draws.

The implementation standardizes the second-stage inputs using the pooled Monte Carlo samples and standardizes the targets $\widehat{\alpha}_i$. The structural model is trained on the scaled variables and transformed back to the original outcome scale for evaluation. Optimization uses AdamW with weight decay 10^{-4}. The reported notebook implementation uses the project routine `CMRLossUStat`, which implements the two sample U-statistic CMR loss in Section 5.3.2.

Synthetic experiments

The synthetic designs are those described in Section 5.5.

Additional details for design 3. The third design uses a heterogeneous mixture first stage. Specifically,

$$W, Z \sim N(0, I_2), \qquad U \sim N(0, 1),$$

and

$$C \mid Z, W \sim \text{Bernoulli}\left(\frac{1}{1+\exp\{-(0.8Z_1 + 0.6W_1 - 0.5W_2)\}}\right).$$

The context-dependent instrument strengths are

$$\pi_1(W) = 0.6 + 0.35\tanh(W_1), \qquad \pi_2(W) = 0.5 + 0.25\mathbf{1}\{W_2 > 0\}.$$

With independent $\xi_1, \xi_2 \sim N(0, 0.25^2)$,

$$X_1 = C(1.1\pi_1(W)Z_1 + 0.5W_1 + 0.7U + \xi_1) + (1-C)(-0.9\pi_1(W)Z_1 + 0.2W_2 + 0.7U + \xi_1),$$

and

$$X_2 = C(0.9\pi_2(W)Z_2 - 0.3W_2 + 0.7U + \xi_2) + (1-C)(-0.7\pi_2(W)Z_2 + 0.4W_1 + 0.7U + \xi_2).$$

The outcome is

$$Y = g_3(X, W) + 0.7U + 0.3\varepsilon, \qquad \varepsilon \sim N(0,1),$$

with g_3 as defined in (5.5.7).

Benchmarks The benchmarks are polynomial OLS, linear IV, sieve 2SLS, and kernel NPIV. Polynomial OLS uses polynomial features of (X, W). Linear IV uses instruments (Z, W). Sieve 2SLS uses polynomial bases in (X, W) and (Z, W), with instrument degrees $b = 1$ and $b = 10$. Kernel NPIV is implemented with the `np` package in R using Tikhonov regularization where computationally feasible.

Structural recovery error is computed on deterministic grids over the interior support. Design 1 uses a grid over X; Design 2 uses a grid over (X_1, X_2); Design 3 averages slice-level errors across representative covariate profiles and treatment-coordinate slices. The Monte Carlo table reports mean error across replications.

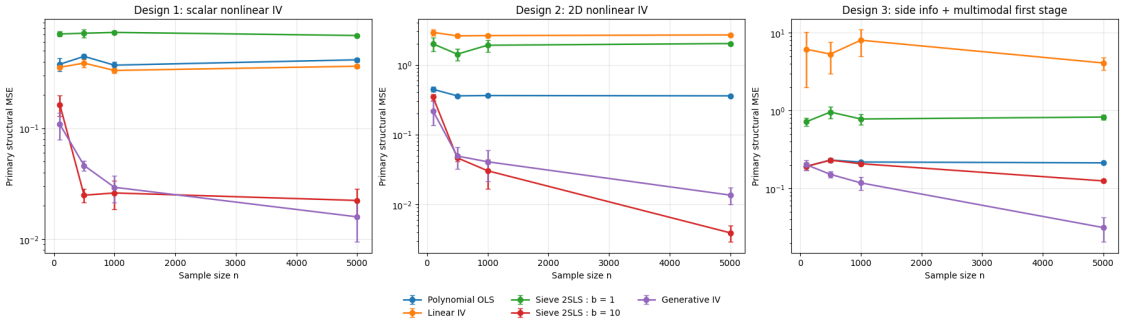


Figure B.1: Performance across designs as sample size, n, increases.

Schooling application

The applied notebook uses the `SchoolingReturns` data from `causalpy`. The outcome is log wage, the endogenous regressor is education, and the excluded instrument is college proximity. Controls are experience, experience squared, urban residence, Southern region, and an indicator for race. After dropping missing values, the sample contains 3010 observations.

OLS and 2SLS provide the linear benchmarks. The generative estimator fits

$$(\log wage, education) \mid (nearcollege, experience, experience^2, smsa, south, Black).$$

The nonlinear second stage uses a degree-five polynomial network in education and controls. Structural schooling curves are evaluated over the empirical interior of the schooling distribution and averaged over the empirical distribution of controls. Marginal returns are computed by numerical differentiation of the fitted structural curve. Heterogeneity figures are obtained by evaluating the fitted structural function and its derivative at fixed covariate profiles.

APPENDIX C

APPENDICES TO CHAPTER 7

C.1 Related Literature

A recent line of research has adopted a causal perspective to formalize methods of statistical reasoning for shifting environments. Notably, the works identify that the invariance principle of causal relations can be leveraged to generalize across arbitrary interventions—a strong notion of robustness, see Scholkopf et al. [2012], Magliacane et al. [2018] and Christiansen et al. [2022]. The ontological relationship between a response and its direct causes is predictive and invariant, and can be shown to be optimal against arbitrarily strong interventions [Peters et al., 2016, Rojas-Carulla et al., 2018]. An influential sequence of papers [Peters et al., 2016, Heinze-Deml et al., 2017, Pfister et al., 2019] casts distribution shifts as the action of some intervention. It compares *labeled data sets* across distinct experimental settings to identify these invariant causal conditionals. However, facing observational data, not all direct causes are observed or identifiable. The interventions might not be diverse or strong enough, which makes causal regression overly conservative. In reality, practitioners face the fundamental tradeoffs between causality and predictability. Rothenhäusler et al. [2021] proposed an actionable methodology to probe the tradeoffs, combining instrumental variable regression (for causality) and ordinary least squares (for predictability), when *access to certain exogenous variables* acting as instruments is available. While Rothenhäusler et al. [2021] has similar high-level commonalities with our work in terms of invariance and predictability tradeoffs, a key distinction is that Rothenhäusler et al. [2021] requires known instruments for identification, whereas we postulate the existence of low-dimensional invariant features that are not necessarily known.

In contrast, the current chapter tackles the domain adaptation problem in observational data when (i) labels are unavailable in the target environment, (ii) unobserved confounding factors drive environment shifts, and (iii) there is no access to exogenous, instrumental

variables to construct quasi-experiments for causal identification [Card and Krueger, 1994, Angrist and Imbens, 1995]. This is a setting less studied in the literature. Our setting inherits the quintessential difficulty of domain adaptation for observational data, the unobserved confounding, and further delineates its role lurking in the environment. In spirit, our work follows and builds upon the literature on improving out-of-domain generalization in ML with causal insights without experimental or quasi-experimental data.

In the learning-theoretic literature, out-of-domain generalization bounds show that provided a stable representation is found under which probability distributions on source and target domain are similar, a hypothesis with low source risk will perform well on target data [Ben-David et al., 2006, Blitzer et al., 2007, Mansour et al., 2009, Ben-David et al., 2010]. This line of work also pioneered the tradeoffs between representation stability and predictability: stable representations may not harness the maximum predictive power. While methodologies that simultaneously search for predictive and stable representations of data [Blitzer et al., 2006, Tzeng et al., 2014, Long et al., 2015, Arjovsky et al., 2019] are now ubiquitous, it is unclear under what data generating mechanisms they perform well. Furthermore, whether the practical optimization method obtains provable guarantees for the learned representation remains to be better understood [Arjovsky et al., 2019].

We anchor the domain adaptation problem in an accessible linear SCM and delineate situations where invariance yields improvement. The methodology in this chapter can be interpreted as a bare-bones instantiation of stability-regularized risk minimization [Tzeng et al., 2014, Ganin et al., 2016, Shen et al., 2018] in the linear SCM. The concrete relationship between stability and predictability will be teased out. We also establish provable characterizations for the learned linear subspace, a concrete step toward representation learning. As we shall see, two notions of stability/invariance unite under the linear SCM: optimizing over a *stable representation* in covariate shifts will align well with searching for the *causal, invariant relationship*, stable across environment shifts.

The problem of covariate shifts, with the conditional concept $Y|X$ unchanged, has spurred

considerable interest in statistics and ML/AI. The statistical study of covariate shift under parametric models dates back to the pathbreaking work of Shimodaira [2000]. Shimodaira [2000] established the asymptotic optimality of vanilla Maximum Likelihood Estimation (MLE) in the well-specified setting and that of weighted MLE under misspecification. Recent non-asymptotic results bolster this result [Ge et al., 2023], and find that the weighted MLE can be minimax optimal under misspecification. The above works operate under a mild covariate shift setting—namely, the likelihood ratio between source and target is often required to be bounded and estimable based on data. Several works explore different facets of covariate shifts specific to the linear setting. Lei et al. [2021] considers the minimax optimal estimator for linear regression under fixed design, where the learner can access some amount of unlabeled target data. However, in the presence of concept shift, near minimax optimal estimation requires target labels. Kalan et al. [2020] provides lower bounds for out-of-distribution generalization in linear and one-hidden-layer neural network models. For hard covariate shifts with an unbounded likelihood ratio, Liang [2024] proposed to study adversarial covariate shifts to understand what extrapolation region adversarial covariate shifts will focus on for a given linear model. Liang [2024] derived a curious dichotomy: depending on the regression or the classification setting, adversarial covariate shift can either be a blessing or a curse to the subsequent learning. In comparison, the current work anchors the problem in a linear SCM where confounding generates both concept and covariate shifts—diverging from the well-specified settings explored in the literature. Our method does not require the bounded likelihood ratio assumption nor the need to estimate such a ratio. Specific to the linear SCM, our work addresses when and how access to unlabeled target samples can boost target domain performance in the presence of concept shifts.

C.2 Discussion

C.2.1 Comparison to the Instrumental Variable Setting

Reflecting on our model (see Figure 7.1), similarities with the classic IV setting are immediate. Our invariant variable, Z, plays a similar role to an instrumental variable, in the sense that it is independent of the confounding environment influence, E. Our confounding variable, E, affects our observed covariates, X, through a restricted linear subspace. This assumption is not necessary in the IV setting. However, unlike the IV literature, we are in the observational setting where Z and E are not known, and in the absence of experiments, cannot necessarily be identified. The linear subspace assumption ensures we can leverage some low-dimensional, stable projection for our goal, which, unlike classical IV, is minimizing target domain risk.

C.2.2 Testability of Assumptions

While it is not necessary (or possible in general) to identify Z or E, our method does hinge on the assumption that covariate shift exists. As the proposed method is based on linear predictors, the covariate shift, if it exists, should be detected from the second moment matrices $\Sigma_{\mathcal{E}} = \mathbb{E}_{\mathcal{E}}[XX^\top]$ for $\mathcal{E} \in \{\mathcal{S}, \mathcal{T}\}$. Hence, using empirical data, we could test for the existence of covariate shifts due to the environmental shift. Suppose we have two sets of i.i.d. covariate vectors from the source and target environments; for simplicity, denote them together as $x_1, \ldots, x_{n+m}$, where the first n of them, $x_1, \ldots, x_n$, are from the source, while the others, $x_{n+1}, \ldots, x_{n+m}$, are from the target. We compute the empirical second moment matrices:

$$\hat{\Sigma}_{\mathcal{S}} := \frac{1}{n} \sum_{i=1}^{n} x_i x_i^\top \quad \text{and} \quad \hat{\Sigma}_{\mathcal{T}} := \frac{1}{m} \sum_{i=n+1}^{n+m} x_i x_i^\top .$$

We compute a suitable statistic comparing $\hat{\Sigma}_{\mathcal{S}}, \hat{\Sigma}_{\mathcal{T}}$, for instance, the operator norm of their difference $\|\hat{\Sigma}_{\mathcal{S}} - \hat{\Sigma}_{\mathcal{T}}\|_{\text{op}}$, which we can compare to its permuted counterpart $\|\hat{\Sigma}^{\sigma}_{\mathcal{S}} - \hat{\Sigma}^{\sigma}_{\mathcal{T}}\|_{\text{op}}$,

where σ is any permutation on $\{1, \ldots, n+m\}$ and

$$\hat{\Sigma}^{\sigma}_{\mathcal{S}} := \frac{1}{n}\sum_{i=1}^{n} x_{\sigma(i)} x_{\sigma(i)}^{\top} \quad \text{and} \quad \hat{\Sigma}^{\sigma}_{\mathcal{T}} := \frac{1}{m}\sum_{i=n+1}^{n+m} x_{\sigma(i)} x_{\sigma(i)}^{\top}.$$

In practice, we sample B permutations $\sigma_1, \ldots, \sigma_B$ from the set of all permutations on $\{1, \ldots, n+m\}$ and see if $\|\hat{\Sigma}_{\mathcal{S}} - \hat{\Sigma}_{\mathcal{T}}\|_{\text{op}}$ is larger than some quantile among $\{\|\hat{\Sigma}^{\sigma_b}_{\mathcal{S}} - \hat{\Sigma}^{\sigma_b}_{\mathcal{T}}\|_{\text{op}}\}_{b=1}^{B}$. It is not necessary to detect which subspace contains this environment shift. We seek a low-dimensional projection best suited for target domain risk—this is not necessarily the space orthogonal to the shifting subspace.

More broadly, a growing literature studies the testing and detection of distribution shifts, including methods designed for weak distributional changes [Hur and Liang, 2025].

C.3 Technical Proofs

C.3.1 Proofs for Section 7.2

Proof of Proposition 7.2.2. Under Assumption 7.2.1, since $\tau > 0$ implies that $\mathbb{E}_{\mathcal{E}}[XX^{\top}]$ is invertible, we can deduce that $\beta_{\mathcal{E}} = (\mathbb{E}_{\mathcal{E}}[XX^{\top}])^{-1}\mathbb{E}_{\mathcal{E}}[XY]$ is uniquely defined. One can verify that

$$\mathbb{E}_{\mathcal{E}}[XY] = \mathbb{E}_{\mathcal{E}}[XX^{\top}]\beta^{\star} + \Delta\mathbb{E}_{\mathcal{E}}[EE^{\top}]\gamma + \Theta\mathbb{E}[Z](\mathbb{E}_{\mathcal{E}}[E])^{\top}\gamma = \mathbb{E}_{\mathcal{E}}[XX^{\top}]\beta^{\star} + \Delta\mathbb{E}_{\mathcal{E}}[EE^{\top}]\gamma\,. \tag{C.3.1}$$

Hence, we have

$$\beta_{\mathcal{E}} = (\mathbb{E}_{\mathcal{E}}[XX^{\top}])^{-1}\mathbb{E}_{\mathcal{E}}[XY] = \beta^{\star} + (\mathbb{E}_{\mathcal{E}}[XX^{\top}])^{-1}\Delta\mathbb{E}_{\mathcal{E}}[EE^{\top}]\gamma\,.$$

Under Assumption 7.2.1, note that

$$\mathbb{E}_{\mathcal{E}}[XX^\top] = \begin{bmatrix}\Theta & \Delta\end{bmatrix}\begin{bmatrix}\mathbb{E}[ZZ^\top] & 0\\ 0 & \mathbb{E}_{\mathcal{E}}[EE^\top]\end{bmatrix}\begin{bmatrix}\Theta^\top\\ \Delta^\top\end{bmatrix} + \tau^2 I_d\,. \tag{C.3.2}$$

As $[\Theta, \Delta] \in \mathrm{St}(d, k+r)$, let $\Gamma \in O_{d-(k+r)}$ whose column space spans the orthogonal complement of $[\Theta, \Delta]$. Then, from (C.3.2), we have

$$\mathbb{E}_{\mathcal{E}}[XX^\top] = \begin{bmatrix}\Theta & \Delta & \Gamma\end{bmatrix}\begin{bmatrix}\mathbb{E}[ZZ^\top] + \tau^2 I_k & 0 & 0\\ 0 & \mathbb{E}_{\mathcal{E}}[EE^\top] + \tau^2 I_r & 0\\ 0 & 0 & \tau^2 I_{d-(k+r)}\end{bmatrix}\begin{bmatrix}\Theta^\top\\ \Delta^\top\\ \Gamma^\top\end{bmatrix}\,, \tag{C.3.3}$$

where we use $\Gamma\Gamma^\top = I_d - \Theta\Theta^\top - \Delta\Delta^\top$. Therefore,

$$\begin{aligned}&(\mathbb{E}_{\mathcal{E}}[XX^\top])^{-1}\\ &= \begin{bmatrix}\Theta & \Delta & \Gamma\end{bmatrix}\begin{bmatrix}(\mathbb{E}[ZZ^\top] + \tau^2 I_k)^{-1} & 0 & 0\\ 0 & (\mathbb{E}_{\mathcal{E}}[EE^\top] + \tau^2 I_r)^{-1} & 0\\ 0 & 0 & (\tau^2 I_{d-(k+r)})^{-1}\end{bmatrix}\begin{bmatrix}\Theta^\top\\ \Delta^\top\\ \Gamma^\top\end{bmatrix}\,.\end{aligned} \tag{C.3.4}$$

Therefore, we have

$$\begin{aligned}\beta_{\mathcal{E}} &= \beta^\star + (\mathbb{E}_{\mathcal{E}}[XX^\top])^{-1}\Delta\mathbb{E}_{\mathcal{E}}[EE^\top]\gamma\\ &= \beta^\star + \Delta\left(\mathbb{E}_{\mathcal{E}}[EE^\top] + \tau^2 I_r\right)^{-1}\mathbb{E}_{\mathcal{E}}[EE^\top]\gamma\\ &= \beta^\star + \Delta\gamma - \tau^2\Delta\left(\mathbb{E}_{\mathcal{E}}[EE^\top] + \tau^2 I_r\right)^{-1}\gamma\,,\end{aligned} \tag{C.3.5}$$

where the second equality uses $(\mathbb{E}_{\mathcal{E}}[XX])^{-1}\Delta = \Delta(\mathbb{E}_{\mathcal{E}}[EE^\top] + \tau^2 I_r)^{-1}$ deduced from (C.3.4). Now, if $\mathbb{E}_{\mathcal{S}}[EE^\top] \neq \mathbb{E}_{\mathcal{T}}[EE^\top]$, there must exist a nonzero coefficient γ such that

$$\tau^2\Delta\left(\mathbb{E}_{\mathcal{S}}[EE^\top] + \tau^2 I_r\right)^{-1}\gamma \neq \tau^2\Delta\left(\mathbb{E}_{\mathcal{T}}[EE^\top] + \tau^2 I_r\right)^{-1}\gamma\,,$$

which implies that $\beta_{\mathcal{S}} \neq \beta_{\mathcal{T}}$. □

Proof of Proposition 7.2.4. Note that $\alpha_{\mathcal{E}}^{\Theta} = (\Theta^\top \mathbb{E}_{\mathcal{E}}[XX^\top]\Theta)^{-1}\Theta^\top \mathbb{E}_{\mathcal{E}}[XY]$. From (C.3.1), we have $\Theta^\top \mathbb{E}_{\mathcal{E}}[XY] = \Theta^\top \mathbb{E}_{\mathcal{E}}[XX^\top]\beta^\star$. From (C.3.3), we have $\Theta^\top \mathbb{E}_{\mathcal{E}}[XX^\top] = (\mathbb{E}[ZZ^\top] + \tau^2 I_k)\Theta^\top$. Therefore,

$$\alpha_{\mathcal{E}}^{\Theta} = (\mathbb{E}[ZZ^\top] + \tau^2 I_k)^{-1}(\mathbb{E}[ZZ^\top] + \tau^2 I_k)\Theta^\top \beta^\star = \Theta^\top \beta^\star \,.$$

As a result, $\beta_{\mathcal{E}}^{\Theta} = \Theta\Theta^\top \beta^\star$, meaning that $\beta_{\mathcal{S}}^{\Theta} = \beta_{\mathcal{T}}^{\Theta}$ always holds regardless of γ. □

Proof of Proposition 7.2.6. Observe that

$$R_{\mathcal{T}}(\beta) = \|\beta_{\mathcal{T}} - \beta\|^2_{\mathbb{E}_{\mathcal{T}}[XX^\top]} + \mathbb{E}_{\mathcal{T}}[Y^2] - \|\beta_{\mathcal{T}}\|^2_{\mathbb{E}_{\mathcal{T}}[XX^\top]} \,.$$

Hence, $R_{\mathcal{T}}(\beta_{\mathcal{S}}^{\Theta}) < R_{\mathcal{T}}(\beta_{\mathcal{S}})$ is equivalent to $\|\beta_{\mathcal{T}} - \beta_{\mathcal{S}}^{\Theta}\|^2_{\mathbb{E}_{\mathcal{T}}[XX^\top]} < \|\beta_{\mathcal{T}} - \beta_{\mathcal{S}}\|^2_{\mathbb{E}_{\mathcal{T}}[XX^\top]}$. From (C.3.5) and $\beta_{\mathcal{S}}^{\Theta} = \Theta\Theta^\top \beta^\star$, we have

$$\begin{aligned}
\beta_{\mathcal{T}} - \beta_{\mathcal{S}}^{\Theta} &= (I_d - \Theta\Theta^\top)\beta^\star + \Delta(\tau^2 I_r + \Lambda_{\mathcal{T}})^{-1}\Lambda_{\mathcal{T}}\gamma = \Delta\left[\Delta^\top \beta^\star + (\tau^2 I_r + \Lambda_{\mathcal{T}})^{-1}\Lambda_{\mathcal{T}}\gamma\right] \,, \\
\beta_{\mathcal{T}} - \beta_{\mathcal{S}} &= \Delta\left[(\tau^2 I_r + \Lambda_{\mathcal{T}})^{-1}\Lambda_{\mathcal{T}}\gamma - (\tau^2 I_r + \Lambda_{\mathcal{S}})^{-1}\Lambda_{\mathcal{S}}\gamma\right] \,,
\end{aligned}$$

where the first equation uses $I_d = \Theta\Theta^\top + \Delta\Delta^\top$. From (C.3.3), we can derive the following and thus complete the proof,

$$\begin{aligned}
\|\beta_{\mathcal{T}} - \beta_{\mathcal{S}}^{\Theta}\|^2_{\mathbb{E}_{\mathcal{T}}[XX^\top]} &= \|\Delta^\top \beta^\star + (\tau^2 I_r + \Lambda_{\mathcal{T}})^{-1}\Lambda_{\mathcal{T}}\gamma\|^2_{\tau^2 I_r + \Lambda_{\mathcal{T}}} \,, \\
\|\beta_{\mathcal{T}} - \beta_{\mathcal{S}}\|^2_{\mathbb{E}_{\mathcal{T}}[XX^\top]} &= \|(\tau^2 I_r + \Lambda_{\mathcal{S}})^{-1}\Lambda_{\mathcal{S}}\gamma - (\tau^2 I_r + \Lambda_{\mathcal{T}})^{-1}\Lambda_{\mathcal{T}}\gamma\|^2_{\tau^2 I_r + \Lambda_{\mathcal{T}}} \,.
\end{aligned}$$

□

C.3.2 Proofs for Section 7.3

Proof of Proposition 7.3.1. By definition of $R_{\mathcal{E}}$, one can deduce that

$$R_{\mathcal{T}}(\beta) - R_{\mathcal{S}}(\beta) = \langle \beta, (\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})\beta \rangle - 2\langle \mathbb{E}_{\mathcal{T}}[XY] - \mathbb{E}_{\mathcal{S}}[XY], \beta \rangle + \mathbb{E}_{\mathcal{T}}[Y^2] - \mathbb{E}_{\mathcal{S}}[Y^2] \ . \quad \text{(C.3.6)}$$

From (C.3.1), we have

$$\mathbb{E}_{\mathcal{T}}[XY] - \mathbb{E}_{\mathcal{S}}[XY] = (\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})\beta^\star + \Delta(\Lambda_{\mathcal{T}} - \Lambda_{\mathcal{S}})\gamma = (\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})(\beta^\star + \Delta\gamma) \ , \quad \text{(C.3.7)}$$

where the second equality uses $\Delta^\top \Delta = I_r$ from Assumption 7.2.1 and

$$\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}} = \Delta(\Lambda_{\mathcal{T}} - \Lambda_{\mathcal{S}})\Delta^\top \quad \text{(C.3.8)}$$

from (C.3.2). Meanwhile, from (7.1.1) and $\mathbb{E}[U] = 0$, we have

$$\mathbb{E}_{\mathcal{E}}[Y^2] = \langle \beta^\star, \Sigma_{\mathcal{E}}\beta^* \rangle + \langle \gamma, \Lambda_{\mathcal{E}}\gamma \rangle + 2\langle \beta^\star, \mathbb{E}_{\mathcal{E}}[XE^\top]\gamma \rangle + \mathbb{E}[U^2] \ .$$

Therefore,

$$\begin{aligned}
\mathbb{E}_{\mathcal{T}}[Y^2] - \mathbb{E}_{\mathcal{S}}[Y^2] &= \langle \beta^\star, (\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})\beta^\star \rangle + \langle \gamma, (\Lambda_{\mathcal{T}} - \Lambda_{\mathcal{S}})\gamma \rangle \\
&\quad + 2\left\langle \beta^\star, \left(\mathbb{E}_{\mathcal{T}}[XE^\top] - \mathbb{E}_{\mathcal{S}}[XE^\top]\right)\gamma \right\rangle \\
&= \langle \beta^\star, (\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})\beta^\star \rangle + \langle \gamma, (\Lambda_{\mathcal{T}} - \Lambda_{\mathcal{S}})\gamma \rangle + 2\langle \beta^\star, \Delta(\Lambda_{\mathcal{T}} - \Lambda_{\mathcal{S}})\gamma \rangle \quad \text{(C.3.9)} \\
&= \langle \beta^\star, (\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})\beta^\star \rangle + \langle \Delta\gamma, (\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})\Delta\gamma \rangle \\
&\quad + 2\langle \beta^\star, (\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})\Delta\gamma \rangle.
\end{aligned}$$

where the second equality is due to (7.1.2) and Assumption 7.2.1 and the third equality uses $\Delta^\top \Delta = I_r$ and (C.3.8) again. Combining (C.3.6), (C.3.7), and (C.3.9), we have (7.3.1). □

Proof of Proposition 7.3.3. By Proposition 7.3.1 and Young's inequality, $\forall \xi > 0$,

$$0 \leq R_{\mathcal{T}}(\beta) - R_{\mathcal{S}}(\beta) \leq (1+\xi)\langle \beta, D\beta \rangle + (1+\xi^{-1})\langle \beta^\star + \Delta\gamma, D(\beta^\star + \Delta\gamma)\rangle \ ,$$

where the second term on the right-hand side does not depend on β. To bound the first term, plug in $\beta = V\alpha$

$$\langle \beta, D\beta \rangle = \langle \alpha\alpha^\top, V^\top DV \rangle \leq \|\alpha\|^2 \|V^\top DV\|_{\mathrm{F}} \ .$$

Apply Young's inequality again, $\forall \zeta > 0$,

$$\|\alpha\|^2 \|V^\top DV\|_{\mathrm{F}} \leq \frac{\zeta}{2}\|\alpha\|^4 + \frac{1}{2\zeta}\|V^\top DV\|_{\mathrm{F}}^2 \ .$$

The proof follows. □

Proof of Proposition 7.3.4. Let $\xi_V = \text{grad } \Phi_{\upsilon,\eta}(V)$. As the Riemannian gradient is equal to the projection of the usual gradient, we have

$$\text{grad } \bar{\Phi}_{\upsilon,\eta}(V) = P_V(\text{grad } \Phi_{\upsilon,\eta}(V)) = P_V(\xi_V) \ ,$$

where

$$P_V(\xi) = (I_d - VV^\top)\xi + V\text{skew}(V^\top \xi) \in T_V \text{St}(d,\ell) \ . \tag{C.3.10}$$

We calculate ξ_V with an application of the Envelope Theorem, see, for example, Milgrom and Segal [2002]. Indeed, the inner minimization, $\min_{\alpha \in \mathbb{R}^\ell} F_{\upsilon,\eta}(V,\alpha)$, admits a unique minimizer

$\alpha_V := (V^\top \Sigma_{\mathcal{S}} V + \upsilon I_\ell)^{-1} V^\top \mathbb{E}_{\mathcal{S}}[XY]$. And so,

$$\begin{aligned}\xi_V &= \text{grad } \Phi_{\upsilon,\eta}(V) = \text{grad } F_{\upsilon,\eta}(V,\alpha)\big|_{\alpha=\alpha_V} \\ &= \left(\Sigma_{\mathcal{S}} V \alpha_V - \mathbb{E}_{\mathcal{S}}[XY]\right) \alpha_V^\top + \eta D V V^\top D V \ .\end{aligned}$$

We evaluate skew$(V^\top \xi)$. Note that

$$\begin{aligned}V^\top \xi &= \left(V^\top \Sigma_{\mathcal{S}} V \alpha_V - V^\top \mathbb{E}_{\mathcal{S}}[XY]\right) \alpha_V^\top + \eta V^\top D V V^\top D V \\ &= -\upsilon \alpha_V \alpha_V^\top + \eta V^\top D V V^\top D V\end{aligned}$$

is a symmetric matrix, thus skew$(V^\top \xi) = 0$. Therefore:

$$\text{grad } \bar{\Phi}_{\upsilon,\eta}(V) = P_V(\xi_V) = (I_d - V V^\top) \left((\Sigma_{\mathcal{S}} V \alpha_V - \mathbb{E}_{\mathcal{S}}[XY]) \alpha_V^\top + \eta D V V^\top D V \right)$$

as required. □

C.3.3 Proofs for Section 7.4

Lemma C.3.1. *For any positive semi-definite matrix $\Sigma \in \mathbb{R}^{d\times d}$ and $V \in \text{St}(d,\ell)$, any $\upsilon > 0$, the following inequality holds in Loewner ordering,*

$$\begin{aligned}\tfrac{\upsilon}{\upsilon+\lambda_{\max}(\Sigma)} \cdot I_d \preceq I_d - \Sigma^{1/2} V (V^\top \Sigma V + \upsilon I_\ell)^{-1} V^\top \Sigma^{1/2} \preceq I_d \ , \\ \Sigma^{1/2} V (V^\top \Sigma V + \upsilon I_\ell)^{-2} V^\top \Sigma^{1/2} \preceq \tfrac{1}{4\upsilon} \cdot I_d \ .\end{aligned}$$

Proof of Lemma C.3.1. The proof uses singular value decomposition (SVD) of $\Sigma^{1/2} V$. Let $\Sigma^{1/2} V = A D B^\top$, $A \in \mathbb{R}^{d\times d}, B \in \mathbb{R}^{\ell\times\ell}$, $D \in \mathbb{R}^{d\times\ell}$, $A^\top A = I_d$, $B B^\top = I_\ell$ and D is

rectangular diagonal. We tackle the first inequality. Note

$$\begin{aligned} I_d - \Sigma^{1/2} V (V^\top \Sigma V + \upsilon I_\ell)^{-1} V^\top \Sigma^{1/2} &= I_d - ADB^\top (BD^\top DB^\top + \upsilon I_\ell)^{-1} BD^\top A^\top \\ &= A(I_d - D(D^\top D + \upsilon I_\ell)^{-1} D^\top) A^\top \ , \end{aligned}$$

where the second equality uses the Woodbury matrix identity. Here, $I_d - D(D^\top D + \upsilon I_\ell)^{-1} D^\top$ is a diagonal matrix with entries

$$(I_d - D(D^\top D + \upsilon I_\ell)^{-1} D^\top)_{ii} = \begin{cases} \frac{\upsilon}{\upsilon + D_{ii}^2} & \text{for } i \leq \ell \ , \\ 1 & \text{otherwise} \ . \end{cases}$$

Note that $D_{ii}^2 = \lambda_i(V^\top \Sigma V)$ where λ_i denotes the i-th largest eigenvalue. We can verify $\lambda_i(V^\top \Sigma V) \leq \lambda_{\max}(\Sigma)$ as $V \in \mathrm{St}(d, \ell)$, thus we arrive at $\frac{\upsilon}{\upsilon + \lambda_{\max}(\Sigma)} \leq \frac{\upsilon}{\upsilon + D_{ii}^2} \leq 1$. The second inequality is established in a similar fashion. Reusing the SVD, we have

$$\Sigma^{1/2} V (V^\top \Sigma V + \upsilon I_\ell)^{-2} V^\top \Sigma^{1/2} = AD(D^\top D + \upsilon I_\ell)^{-2} D^\top A^\top \ .$$

The entries of the inner diagonal matrix read,

$$(D(D^\top D + \upsilon I_\ell)^{-2} D^\top)_{ii} = \begin{cases} \frac{D_{ii}^2}{(D_{ii}^2 + \upsilon)^2} & \text{for } i \leq \ell \ , \\ 0 & \text{otherwise} \ . \end{cases}$$

Then, $\frac{D_{ii}^2}{(D_{ii}^2+\upsilon)^2} \leq \max_{d>0} \frac{d^2}{(d^2+\upsilon)^2} = \frac{1}{4\upsilon}$. □

Proof of Theorem 7.4.1. Let $M := \Lambda_{\mathcal{T}} - \Lambda_{\mathcal{S}}$. Under Assumption 7.3.2, M is a positive definite matrix. The first-order condition of the Stiefel manifold optimization (7.3.6) reads

$$(I_d - VV^\top)(\mathbb{E}_{\mathcal{S}}[XY] - \mathbb{E}_{\mathcal{S}}[XX^\top] V \alpha_V) \alpha_V^\top = \eta (I_d - VV^\top) DVV^\top DV \ . \tag{C.3.11}$$

Define the canonical angles $A := V^\top \Delta \in \mathbb{R}^{\ell \times r}$. Taking the Frobenius norm of the right-hand side of (C.3.11), we have

$$\begin{aligned}
&\eta^2 \mathrm{Tr}\big[V^\top D V V^\top D (I_d - VV^\top) D V V^\top D V\big] \\
&= \eta^2 \mathrm{Tr}\big[AMA^\top AM(I_r - A^\top A) MA^\top AMA^\top\big] \\
&\geq \eta^2 \lambda_{\min}(I_r - A^\top A) \cdot \lambda_{\min}(M) \cdot \mathrm{Tr}\big[AMA^\top AMA^\top AMA^\top\big] \\
&\geq \eta^2 \lambda_{\min}(I_r - A^\top A) \cdot \lambda_{\min}(M)^4 \cdot \|AA^\top\|_{\mathrm{op}}^3 \,.
\end{aligned}$$

In the inequalities above, we use the fact that if $B \succeq C$, then $DBD^\top \succeq DCD^\top$, in Loewner ordering. The last inequality leverages the additional facts that $\mathrm{Tr}\big[(AMA^\top)^3\big] \geq \|AMA^\top\|_{\mathrm{op}}^3$ and $\|AMA^\top\|_{\mathrm{op}} \geq \lambda_{\min}(M)\|AA^\top\|_{\mathrm{op}}$.

The Frobenius norm on the left-hand side of (C.3.11) is

$$\|(I_d - VV^\top)(\mathbb{E}_S[XY] - \mathbb{E}_S[XX^\top] V \alpha_V)\|^2 \cdot \|\alpha_V\|^2 \,. \tag{C.3.12}$$

Recall the two facts in Lemma C.3.1,

$$\tfrac{\upsilon}{\upsilon + \lambda_{\max}(\Sigma)} \cdot I_d \preceq I_d - \Sigma^{1/2} V (V^\top \Sigma V + \upsilon I_\ell)^{-1} V^\top \Sigma^{1/2} \preceq I_d \,, \tag{C.3.13}$$

$$\Sigma^{1/2} V (V^\top \Sigma V + \upsilon I_\ell)^{-2} V^\top \Sigma^{1/2} \preceq \tfrac{1}{4\upsilon} \cdot I_d \,. \tag{C.3.14}$$

Now we control each term in (C.3.12).

For the first term in (C.3.12), we have

$$\begin{aligned}
&(I_d - VV^\top)(\mathbb{E}_S[XY] - \mathbb{E}_S[XX^\top] V \alpha_V) \\
&= (I_d - VV^\top)\big(I_d - \mathbb{E}_{\mathcal{S}}[XX^\top] V (V^\top \mathbb{E}_{\mathcal{S}}[XX^\top] V + \upsilon I_\ell)^{-1} V^\top\big) \mathbb{E}_{\mathcal{S}}[XY] \,.
\end{aligned}$$

Therefore, denote $\Sigma := \mathbb{E}_{\mathcal{S}}[XX^\top]$

$$
\begin{aligned}
&\|(I_d - VV^\top)(\mathbb{E}_{\mathcal{S}}[XY] - \mathbb{E}_{\mathcal{S}}[XX^\top]V\alpha_V)\|^2 \\
&\le \|\big(I_d - \mathbb{E}_{\mathcal{S}}[XX^\top]V(V^\top\mathbb{E}_{\mathcal{S}}[XX^\top]V + \upsilon I_\ell)^{-1}V^\top\big)\mathbb{E}_{\mathcal{S}}[XY]\|^2 \\
&= \|\Sigma^{1/2}\big(I_d - \Sigma^{1/2}V(V^\top\Sigma V + \upsilon I_\ell)^{-1}V^\top\Sigma^{1/2}\big)\Sigma^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^2 \\
&\le \lambda_{\max}(\Sigma)\cdot\|\big(I_d - \Sigma^{1/2}V(V^\top\Sigma V + \upsilon I_\ell)^{-1}V^\top\Sigma^{1/2}\big)\Sigma^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^2 \quad \text{by (C.3.13)} \\
&\le \lambda_{\max}(\Sigma)\cdot\|\Sigma^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^2 \,.
\end{aligned}
$$

For the second term in (C.3.12), we know

$$
\begin{aligned}
\|\alpha_V\|^2 &= \|(V^\top\Sigma V + \upsilon I_\ell)^{-1}V^\top\mathbb{E}_{\mathcal{S}}[XY]\|^2 \\
&\le \lambda_{\max}\big(\Sigma^{1/2}V(V^\top\Sigma V + \upsilon I_\ell)^{-2}V^\top\Sigma^{1/2}\big)\cdot\|\Sigma^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^2 \quad \text{by (C.3.14)} \\
&\le \frac{1}{4\upsilon}\|\Sigma^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^2 \,.
\end{aligned}
$$

Putting things together, for $V \in \mathcal{S}_\Delta(\delta)$, we have that the canonical angles $A = V^\top\Delta \in \mathbb{R}^{\ell\times r}$ satisfies $1 - \lambda_{\max}(A^\top A) \ge 1 - \|A\|_{\mathrm{op}}^2 \ge \delta$, and that

$$
\eta^2\delta\lambda_{\min}(M)^4\cdot\|A^\top A\|_{\mathrm{op}}^3 \le \lambda_{\max}(\Sigma_{\mathcal{S}})\cdot\|\Sigma_{\mathcal{S}}^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^2\frac{1}{4\upsilon}\|\Sigma_{\mathcal{S}}^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^2 \,.
$$

Therefore, we have proved

$$
\|V^\top\Delta\|_{\mathrm{op}}^6 \le \frac{\delta^{-1}\lambda_{\max}(\Sigma_{\mathcal{S}})\|\Sigma_{\mathcal{S}}^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^4}{\lambda_{\min}(\Delta^\top D\Delta)^4}\frac{1}{4\upsilon\eta^2} \,.
$$

□

Proof of Theorem 7.4.3. In view of (7.3.1), by Young's inequality, we have

$$
R_{\mathcal{T}}(\beta^{\upsilon,\eta}) - R_{\mathcal{S}}(\beta^{\upsilon,\eta}) \le (1+\epsilon)\langle\beta^\star + \Delta\gamma, D(\beta^\star + \Delta\gamma)\rangle + (1+\epsilon^{-1})\langle V\alpha_V, DV\alpha_V\rangle \,.
$$

Note

$$\begin{aligned}\langle V\alpha_V, DV\alpha_V\rangle &\le \|\alpha_V\|^2 \cdot \|V^\top DV\|_{\mathrm{op}} \\ &\le \|\alpha_V\|^2 \cdot \|V^\top \Delta M \Delta^\top V\|_{\mathrm{op}} \\ &\le \|\alpha_V\|^2 \cdot \lambda_{\max}(M)\|V^\top \Delta\|_{\mathrm{op}}^2 \,.\end{aligned}$$

Recall the fact that

$$\|\alpha_V\|^2 \le \frac{1}{4\upsilon}\|\Sigma^{-1/2}\mathbb{E}_{\mathcal{S}}[XY]\|^2$$

and the bound on $\|V^\top \Delta\|_{\mathrm{op}}^6$ established in Theorem 7.4.1, we finish the proof after separating out the terms independent of η, υ and defining them as $\mathsf{S}_{\epsilon,\delta}$. □

Proof of Theorem 7.4.5. First, we claim

$$\Phi_{\upsilon,\eta}(\widehat{V}) - \min_{V\in \mathrm{St}(d,\ell)} \Phi_{\upsilon,\eta}(V) \le 2 \sup_{V\in \mathrm{St}(d,\ell)} |\widehat{\Phi}_{\upsilon,\eta}(V) - \Phi_{\upsilon,\eta}(V)| \,. \tag{C.3.15}$$

To see this, pick a minimizer $V^\star$ of $\Phi_{\upsilon,\eta}$ over $\mathrm{St}(d,\ell)$. Then, we have

$$\begin{aligned}\Phi_{\upsilon,\eta}(\widehat{V}) - \Phi_{\upsilon,\eta}(V^\star) &= \Phi_{\upsilon,\eta}(\widehat{V}) - \widehat{\Phi}_{\upsilon,\eta}(\widehat{V}) + \widehat{\Phi}_{\upsilon,\eta}(\widehat{V}) - \widehat{\Phi}_{\upsilon,\eta}(V^\star) + \widehat{\Phi}_{\upsilon,\eta}(V^\star) - \Phi_{\upsilon,\eta}(V^\star) \\ &\le \Phi_{\upsilon,\eta}(\widehat{V}) - \widehat{\Phi}_{\upsilon,\eta}(\widehat{V}) + \widehat{\Phi}_{\upsilon,\eta}(V^\star) - \Phi_{\upsilon,\eta}(V^\star) \\ &\le 2 \sup_{V\in \mathrm{St}(d,\ell)} |\widehat{\Phi}_{\upsilon,\eta}(V) - \Phi_{\upsilon,\eta}(V)| \,,\end{aligned}$$

where the first inequality follows as $\widehat{V}$ minimizes $\widehat{\Phi}_{\upsilon,\eta}$.

Next, let $\widehat{R}_{\mathcal{S}}$, $\widehat{\Sigma}_{\mathcal{T}}$, and $\widehat{\Sigma}_{\mathcal{S}}$ be the empirical versions of $R_{\mathcal{S}}$, $\Sigma_{\mathcal{T}}$, and $\Sigma_{\mathcal{S}}$, respectively.

Then, we have

$$2(\widehat{\Phi}_{\upsilon,\eta}(V) - \Phi_{\upsilon,\eta}(V)) = \min_{\alpha\in\mathbb{R}^\ell}\left(\widehat{R}_{\mathcal{S}}(V\alpha) + \upsilon\|\alpha\|^2\right) - \min_{\alpha\in\mathbb{R}^\ell}\left(R_{\mathcal{S}}(V\alpha) + \upsilon\|\alpha\|^2\right) + \frac{\eta}{2}\left(\|V^\top(\widehat{\Sigma}_{\mathcal{T}} - \widehat{\Sigma}_{\mathcal{S}})V\|_{\mathrm{F}}^2 - \|V^\top(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})V\|_{\mathrm{F}}^2\right) . \tag{C.3.16}$$

For any $V \in \mathrm{St}(d,\ell)$, we have $\alpha_V = (V^\top\Sigma_{\mathcal{S}}V + \upsilon I_\ell)^{-1}V^\top\mathbb{E}_{\mathcal{S}}[XY]$, which yields

$$\|\alpha_V\| \le \|(V^\top\Sigma_{\mathcal{S}}V + \upsilon I_\ell)^{-1}\|_{\mathrm{op}} \cdot \|V^\top\mathbb{E}_{\mathcal{S}}[XY]\| \le \frac{M^2}{\|V^\top\Sigma_{\mathcal{S}}V + \upsilon I_\ell\|_{\mathrm{op}}} \le \frac{M^2}{\upsilon} .$$

Similarly, $\|\widehat{\alpha}_V\| \le \frac{M^2}{\upsilon}$ for $\widehat{\alpha}_V := \arg\min_{\alpha\in\mathbb{R}^\ell}\widehat{F}_{\upsilon,\eta}(V,\alpha) = (V^\top\widehat{\Sigma}_{\mathcal{S}}V + \upsilon I_\ell)^{-1}V^\top\widehat{\mathbb{E}}_{\mathcal{S}}[XY]$, where $\widehat{F}_{\upsilon,\eta}, \widehat{\mathbb{E}}_{\mathcal{S}}$ denote the empirical versions of $F_{\upsilon,\eta}, \mathbb{E}_{\mathcal{S}}$, respectively. Therefore,

$$\begin{aligned}
&\left|\min_{\alpha\in\mathbb{R}^\ell}\left(\widehat{R}_{\mathcal{S}}(V\alpha) + \upsilon\|\alpha\|^2\right) - \min_{\alpha\in\mathbb{R}^\ell}\left(R_{\mathcal{S}}(V\alpha) + \upsilon\|\alpha\|^2\right)\right| \\
&= \left|\min_{\substack{\alpha\in\mathbb{R}^\ell \\ \|\alpha\|\le M^2/\upsilon}}\left(\widehat{R}_{\mathcal{S}}(V\alpha) + \upsilon\|\alpha\|^2\right) - \min_{\substack{\alpha\in\mathbb{R}^\ell \\ \|\alpha\|\le M^2/\upsilon}}\left(R_{\mathcal{S}}(V\alpha) + \upsilon\|\alpha\|^2\right)\right| \\
&\le \sup_{\substack{\alpha\in\mathbb{R}^\ell \\ \|\alpha\|\le M^2/\upsilon}}\left|\widehat{R}_{\mathcal{S}}(V\alpha) - R_{\mathcal{S}}(V\alpha)\right| .
\end{aligned}$$

Combining (C.3.15), (C.3.16), and

$$\sup_{V\in\mathrm{St}(d,\ell)}\sup_{\substack{\alpha\in\mathbb{R}^\ell \\ \|\alpha\|\le M^2/\upsilon}}\left|\widehat{R}_{\mathcal{S}}(V\alpha) - R_{\mathcal{S}}(V\alpha)\right| = \sup_{\substack{\beta\in\mathbb{R}^d \\ \|\beta\|\le M^2/\upsilon}}\left|\widehat{R}_{\mathcal{S}}(\beta) - R_{\mathcal{S}}(\beta)\right| ,$$

we have

$$\begin{aligned}
&\Phi_{\upsilon,\eta}(\widehat{V}) - \min_{V\in\mathrm{St}(d,\ell)} \Phi_{\upsilon,\eta}(V) \\
&\quad = \sup_{\substack{\beta\in\mathbb{R}^d \\ \|\beta\|\le M^2/\upsilon}} \left|\widehat{R}_{\mathcal{S}}(\beta) - R_{\mathcal{S}}(\beta)\right| \\
&\qquad + \frac{\eta}{2} \sup_{V\in\mathrm{St}(d,\ell)} \left| \|V^\top(\widehat{\Sigma}_{\mathcal{T}} - \widehat{\Sigma}_{\mathcal{S}})V\|_{\mathrm{F}}^2 - \|V^\top(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})V\|_{\mathrm{F}}^2 \right| .
\end{aligned} \tag{C.3.17}$$

By the symmetrization argument, we have

$$\mathbb{E} \sup_{\substack{\beta\in\mathbb{R}^d \\ \|\beta\|\le M^2/\upsilon}} \left|\widehat{R}_{\mathcal{S}}(\beta) - R_{\mathcal{S}}(\beta)\right| \le \frac{2}{n}\mathbb{E} \sup_{\substack{\beta\in\mathbb{R}^d \\ \|\beta\|\le M^2/\upsilon}} \left|\sum_{i=1}^n \sigma_i (y_i - \langle\beta, x_i\rangle)^2\right| ,$$

where $\sigma_1, \ldots, \sigma_n$ are i.i.d. Rademacher variables that are independent of $\{(x_i, y_i)\}_{i=1}^n$ and $\{\tilde{x}_i\}_{i=1}^m$. As $|y_i - \langle\beta, x_i\rangle| \le M + \frac{M^3}{\upsilon} =: L$ for all i and $\|\beta\| \le M^2/\upsilon$, we can apply the contraction principle of the Rademacher complexity to the $(2L)$-Lipschitz function $\phi(t) = t^2$ on $[-L, L]$, which yields

$$\begin{aligned}
\mathbb{E} \sup_{\substack{\beta\in\mathbb{R}^d \\ \|\beta\|\le M^2/\upsilon}} \left|\widehat{R}_{\mathcal{S}}(\beta) - R_{\mathcal{S}}(\beta)\right| &\le \frac{8(M + \frac{M^3}{\upsilon})}{n} \times \mathbb{E} \sup_{\substack{\beta\in\mathbb{R}^d \\ \|\beta\|\le M^2/\upsilon}} \left|\sum_{i=1}^n \sigma_i (y_i - \langle\beta, x_i\rangle)\right| \\
&\le \frac{8(M + \frac{M^3}{\upsilon})}{n} \left(\mathbb{E}\left|\sum_{i=1}^n \sigma_i y_i\right| + \mathbb{E} \sup_{\substack{\beta\in\mathbb{R}^d \\ \|\beta\|\le M^2/\upsilon}} \left|\sum_{i=1}^n \sigma_i \langle\beta, x_i\rangle\right| \right) ,
\end{aligned}$$

where the second inequality follows from the triangle inequality. We have

$$\mathbb{E}\left|\sum_{i=1}^n \sigma_i y_i\right| \le \sqrt{\mathbb{E}\left|\sum_{i=1}^n \sigma_i y_i\right|^2} = \sqrt{\mathbb{E}\sum_{i=1}^n |Y_i|^2} \le M\sqrt{n} ,$$

where the first inequality is Jensen's inequality, and the equality follows from the indepen-

dence of y_i's and σ_i's. Similarly, we can deduce $\mathbb{E}\,\|\sum_{i=1}^n \sigma_i x_i\| \le M\sqrt{n}$, which yields

$$\mathbb{E} \sup_{\substack{\beta\in\mathbb{R}^d \\ \|\beta\|\le M^2/\upsilon}} \left|\sum_{i=1}^n \sigma_i \langle \beta, x_i\rangle\right| = \frac{M^2}{\upsilon}\mathbb{E}\left\|\sum_{i=1}^n \sigma_i x_i\right\| \le \frac{M^3}{\upsilon}\sqrt{n}\,.$$

Therefore,

$$\mathbb{E} \sup_{\substack{\beta\in\mathbb{R}^d \\ \|\beta\|\le M^2/\upsilon}} \left|\widehat{R}_{\mathcal{S}}(\beta) - R_{\mathcal{S}}(\beta)\right| \le \frac{8(M+\frac{M^3}{\upsilon})^2}{\sqrt{n}}\,.$$

If we change a data point (x_i, y_i) to (x_i', y_i'), then the change in $\widehat{R}_{\mathcal{S}}(\beta)$ for any $\beta \in \mathbb{R}^d$ with $\|\beta\| \le M^2/\upsilon$ can be bounded by

$$\frac{1}{n}|(y_i - \langle\beta, x_i\rangle)^2 - (y_i' - \langle\beta, x_i'\rangle)^2| \le \frac{1}{n}\sup_{\substack{(x,y)\in\mathbb{R}^d\times[-M,M] \\ \|x\|\le M}} (y - \langle\beta, x\rangle)^2 \le \frac{(M+\frac{M^3}{\upsilon})^2}{n}\,.$$

By McDiarmid's inequality, for any $\delta \in (0,1)$, the inequality

$$\sup_{\substack{\beta\in\mathbb{R}^d \\ \|\beta\|\le M^2/\upsilon}} \left|\widehat{R}_{\mathcal{S}}(\beta) - R_{\mathcal{S}}(\beta)\right| \le \mathbb{E} \sup_{\substack{\beta\in\mathbb{R}^d \\ \|\beta\|\le M^2/\upsilon}} \left|\widehat{R}_{\mathcal{S}}(\beta) - R_{\mathcal{S}}(\beta)\right| + (M+\frac{M^3}{\upsilon})^2\sqrt{\frac{\log\frac{1}{\delta}}{2n}}$$

holds with probability at least $1-\delta$. Hence,

$$\sup_{\substack{\beta\in\mathbb{R}^d \\ \|\beta\|\le M^2/\upsilon}} \left|\widehat{R}_{\mathcal{S}}(\beta) - R_{\mathcal{S}}(\beta)\right| \le \frac{8(M+\frac{M^3}{\upsilon})^2}{\sqrt{n}} + (M+\frac{M^3}{\upsilon})^2\sqrt{\frac{\log\frac{1}{\delta}}{2n}} \tag{C.3.18}$$

holds with probability at least $1-\delta$.

Next, for $V \in \mathrm{St}(d,\ell)$, we have

$$\|V^\top(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})V\|_{\mathrm{F}} \le \|V^\top\|_{\mathrm{op}} \times \|(\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}})\|_{\mathrm{op}} \times \|V\|_{\mathrm{F}} = \sqrt{\ell}\cdot\|\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}}\|_{\mathrm{op}}\,.$$

Similarly,

$$\begin{aligned}\|V^\top(\widehat{\Sigma}_{\mathcal{T}}-\widehat{\Sigma}_{\mathcal{S}})V\|_{\mathrm{F}} &\le \sqrt{\ell}(\|\widehat{\Sigma}_{\mathcal{T}}\|_{\mathrm{op}}+\|\widehat{\Sigma}_{\mathcal{S}}\|_{\mathrm{op}})\\ &\le \sqrt{\ell}\left(\frac{1}{n}\sum_{i=1}^{n}\|x_i\|^2+\frac{1}{m}\sum_{i=1}^{m}\|\tilde{x}_i\|^2\right)\\ &\le 2\sqrt{\ell}M^2\ .\end{aligned}$$

Therefore,

$$\begin{aligned}&\sup_{V\in\mathrm{St}(d,\ell)}\left|\|V^\top(\widehat{\Sigma}_{\mathcal{T}}-\widehat{\Sigma}_{\mathcal{S}})V\|_{\mathrm{F}}^2-\|V^\top(\Sigma_{\mathcal{T}}-\Sigma_{\mathcal{S}})V\|_{\mathrm{F}}^2\right|\\ &\quad\le\sqrt{\ell}\left(\|\Sigma_{\mathcal{T}}-\Sigma_{\mathcal{S}}\|_{\mathrm{op}}+2M^2\right)\times\sup_{V\in\mathrm{St}(d,\ell)}\left|\|V^\top(\widehat{\Sigma}_{\mathcal{T}}-\widehat{\Sigma}_{\mathcal{S}})V\|_{\mathrm{F}}-\|V^\top(\Sigma_{\mathcal{T}}-\Sigma_{\mathcal{S}})V\|_{\mathrm{F}}\right|\end{aligned}$$

Meanwhile, for $V\in\mathrm{St}(d,\ell)$, we have

$$\begin{aligned}\left|\|V^\top(\widehat{\Sigma}_{\mathcal{T}}-\widehat{\Sigma}_{\mathcal{S}})V\|_{\mathrm{F}}-\|V^\top(\Sigma_{\mathcal{T}}-\Sigma_{\mathcal{S}})V\|_{\mathrm{F}}\right| &\le \|V^\top(\widehat{\Sigma}_{\mathcal{T}}-\Sigma_{\mathcal{T}})V\|_{\mathrm{F}}+\|V^\top(\widehat{\Sigma}_{\mathcal{S}}-\Sigma_{\mathcal{S}})V\|_{\mathrm{F}}\\ &\le\sqrt{\ell}\left(\|\widehat{\Sigma}_{\mathcal{T}}-\Sigma_{\mathcal{T}}\|_{\mathrm{op}}+\|\widehat{\Sigma}_{\mathcal{S}}-\Sigma_{\mathcal{S}}\|_{\mathrm{op}}\right)\ .\end{aligned}$$

Therefore, we have

$$\begin{aligned}&\sup_{V\in\mathrm{St}(d,\ell)}\left|\|V^\top(\widehat{\Sigma}_{\mathcal{T}}-\widehat{\Sigma}_{\mathcal{S}})V\|_{\mathrm{F}}^2-\|V^\top(\Sigma_{\mathcal{T}}-\Sigma_{\mathcal{S}})V\|_{\mathrm{F}}^2\right|\\ &\quad\le\ell\left(\|\Sigma_{\mathcal{T}}-\Sigma_{\mathcal{S}}\|_{\mathrm{op}}+2M^2\right)\times\left(\|\widehat{\Sigma}_{\mathcal{T}}-\Sigma_{\mathcal{T}}\|_{\mathrm{op}}+\|\widehat{\Sigma}_{\mathcal{S}}-\Sigma_{\mathcal{S}}\|_{\mathrm{op}}\right)\ .\end{aligned}$$

We bound the terms $\|\widehat{\Sigma}_{\mathcal{T}}-\Sigma_{\mathcal{T}}\|_{\mathrm{op}}$ and $\|\widehat{\Sigma}_{\mathcal{S}}-\Sigma_{\mathcal{S}}\|_{\mathrm{op}}$ using the standard covariance estimation results based on the matrix Bernstein inequality. For instance, based on Theorem 1.6.2 and Section 1.6.3 of Tropp et al. [2015], we have for any $t>0$,

$$\mathbb{P}(\|\widehat{\Sigma}_{\mathcal{T}}-\Sigma_{\mathcal{T}}\|_{\mathrm{op}}\ge t)\le 2d\cdot\exp\left(-\frac{t^2/2}{\frac{M^2\|\Sigma_{\mathcal{T}}\|_{\mathrm{op}}}{n}+\frac{2M^2t}{3n}}\right)\ .$$

From this, we can deduce that for any $\delta \in (0,1)$, the inequality

$$\|\widehat{\Sigma}_{\mathcal{T}} - \Sigma_{\mathcal{T}}\|_{\mathrm{op}} < \frac{4M^2 \log \frac{2d}{\delta}}{3n} + \sqrt{\frac{2M^2 \|\Sigma_{\mathcal{T}}\|_{\mathrm{op}} \log \frac{2d}{\delta}}{n}}$$

holds with probability at least $1-\delta$. Similarly, we can bound $\|\widehat{\Sigma}_{\mathcal{S}} - \Sigma_{\mathcal{S}}\|_{\mathrm{op}}$. Combining these results with (C.3.17) and (C.3.18) through the union bound, the inequality

$$\begin{aligned}
&\Phi_{\upsilon,\eta}(\widehat{V}) - \min_{V \in \mathrm{St}(d,\ell)} \Phi_{\upsilon,\eta}(V) \\
&\quad \leq \frac{8(M + \frac{M^3}{\upsilon})^2}{\sqrt{n}} + (M + \frac{M^3}{\upsilon})^2 \sqrt{\frac{\log \frac{3}{\delta}}{2n}} \\
&\qquad + \frac{\eta\ell(\|\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}}\|_{\mathrm{op}} + 2M^2)}{2} \times \left(\frac{8M^2 \log \frac{6d}{\delta}}{3n} + \sqrt{\frac{8M^2 \max(\|\Sigma_{\mathcal{T}}\|_{\mathrm{op}}, \|\Sigma_{\mathcal{S}}\|_{\mathrm{op}}) \log \frac{6d}{\delta}}{n}} \right) \\
&\quad \leq 9(M + \frac{M^3}{\upsilon})^2 \sqrt{\frac{\log \frac{3}{\delta}}{n}} + \mathsf{C}_{M,\Sigma_{\mathcal{T}},\Sigma_{\mathcal{S}}} \cdot \eta\ell \left(\frac{\log \frac{6d}{\delta}}{n} + \sqrt{\frac{\log \frac{6d}{\delta}}{n}} \right)
\end{aligned}$$

holds with probability at least $1-\delta$, where

$$\mathsf{C}_{M,\Sigma_{\mathcal{T}},\Sigma_{\mathcal{S}}} = \frac{\|\Sigma_{\mathcal{T}} - \Sigma_{\mathcal{S}}\|_{\mathrm{op}} + 2M^2}{2} \times \max\left(\frac{8M^2}{3}, \sqrt{8M^2 \max(\|\Sigma_{\mathcal{T}}\|_{\mathrm{op}}, \|\Sigma_{\mathcal{S}}\|_{\mathrm{op}})} \right) .$$

□