The Price of Fairness on Regression Models

November 12, 2025

Motivation

We are interested in characterizing the tradeoffs between fairness and predictive power in regression models. These models try to find the best approximation function $f_\theta : \mathcal{X} \to \mathcal{Y}$, to map the elements of a feature space $\mathcal{X}$ to the elements of an objective space $\mathcal{Y}$. One potential harm of trying to find the best approximation f_θ for a given dataset $\{y_i, x_i\}_{i=1}^n$, is that usually the measure of error that is optimized (e.g. Mean Squared Error), only focuses on the average, and thus this solutions are usually myopic to disparities; for example, the variance among samples, disparities on the error of different demographic rules, etc. We will refer to the presence of these disparities as unfairness in the regression model.

It is expected that, if we observe a certain level of fairness (for a given fairness measure) associated with the minimum-error approximation function, and we try to decrease the unfairness of it, then we would probably get a suboptimal solution in terms of the error objective. In other words, it is expected that there exists a *tradeoff* between the level of fairness and the quality of the approximation. Given this, if the unfairness of our model is not acceptable for our goals (e.g. high level of discrimination of our approximation based on gender, race, etc.), then our model does not really serve any purpose; i.e., a good quality of the approximation does not always guarantee a well-suited model. We will focus then on answering the following questions of interest.

- **Q1:** Can we measure how much approximation error we sacrifice if we want to achieve a certain level of fairness?
- **Q2:** How much of the approximation quality are we willing to sacrifice to achieve a certain level of fairness in our model?

To answer these questions, we need to be able to quantify the exact *tradeoffs* that exist between given measures of fairness and approximation quality. For that, we need to further answer the following.

- **Q3:** Can we derive a closed-form expression for the *tradeoff* between fairness and predictive power of our model?
- **Q4:** If no closed-form expression exists, can we derive good-enough approximations for the *tradeoff* between fairness and predictive power of our model?

1 Introduction

1.1 Regression models in terms of approximation error and fairness

To address our questions, we first need to define measures for approximation error and the level of fairness or unfairness. Let us denote the general measure of approximation error of a given f_θ

as the scalar mapping $J : \Theta \to \mathbb{R}$ (e.g., MSE, MAE, etc.), and, likewise, the general measure of unfairness level as $F : \Theta \to \mathbb{R}$ (e.g., maximum error of a group, difference between maximum and minimum error of two groups, etc.), where Θ defines our hypothesis space that supports all the possible functions f_θ that approximate $y \in \mathcal{Y}$, given our assumptions on the structure of $\mathcal{Y}$. In the general regression approach, one aims to choose the best possible set of parameters $\theta \in \Theta$ to minimize the approximation error $J(\theta)$, . In other words, regression models solve the optimization problem

$$\begin{aligned} (P_0) := \min_{\theta} \quad & J(\theta) \\ s.t. \quad & \theta \in \Theta, \end{aligned} \tag{1}$$

where the solution θ_0 produces f_{θ_0}, the approximation with the least approximation error, $J(\theta_0)$. This problem also gives us the baseline unfairness level $F(\theta_0)$. Any improvement $\delta > 0$ on fairness that we wish to achieve has to produce a lower level of F. This can be achieved by solving the fairness-constrained optimization problem

$$\begin{aligned} (P_\delta) := \min_{\theta} \quad & J(\theta) - J(\theta_0) \\ s.t. \quad & F(\theta) - F(\theta_0) \leq -\delta \ , \\ & \theta \in \Theta \end{aligned} \tag{2}$$

where $J(\theta_0)$ is just a constant for this problem. Note that we could also write the constraint on the fairness with equality, to that exact level of change, but we set the upper bound for now. Here, the *tradeoff* structure is clear, since we are decreasing $F(\theta_0)$ by at least δ, and, because the feasible region of P_δ is a subset of the feasible region of P_0, then $J(\theta_\delta) \geq J(\theta_0)$.

1.2 The price of fairness in terms of approximation error

The problem with trying to set a goal for the fairness level for $F(\theta_\delta)$, and trying to identify the sacrifice of that level in terms of the difference of approximation error $\Delta J_\delta = J(\theta_\delta) - J(\delta_0)$, is that this depends on the magnitude of the measures. For example, solving (P_0) with the error metric J_1 and solving it with the metric $J_2 = 2J_1$ gives us the same solution set for θ_0, but the price that we pay in ΔJ_δ is the double for the latter. We need to define an extra metric that is independent of scaling transformations, so we use an equivalent of the price of fairness discussed by Bertsimas et al. (2011), i.e., we define the Price Of Fairness (POF) obtained via solving P_δ as

$$\text{POF}(\theta_\delta, \theta_0) := \frac{J(\theta_\delta) - J(\theta_0)}{J(\theta_0)}, \tag{3}$$

which is simply the percentual change in J given the fair solution θ_δ, with respect to the minimum approximation error solution θ_0. Following a similar scheme, we can also define the Percentual Fairness Gain (PFG) obtained with θ_δ as

$$\text{PFG}(\theta_\delta, \theta_0) := \frac{F(\theta_0) - F(\theta_\delta)}{F(\theta_0)}, \tag{4}$$

which is, again, the percentual change in F after setting the maximum unfairness level in P_δ. Note that both metrics are nonnegative, and that we can either use the original difference or these new metrics to study the *tradeoff* between approximation error and fairness.

1.3 On the tradeoff between approximation error and fairness

Note that this setting encompasses a broad diversity of models (e.g., linear regression, regression trees, neural network regression, etc.), approximation error measures, and also fairness measures. Each model, approximation error measure, and fairness measure behaves differently, and the interactions between them are even more diverse. Therefore, we don't expect to define a universal method to derive the *tradeoff* between our two measure, but instead define schemes that can help answer this questions in the most common and/or most relevant settings.

In general, to be able to answer **Q1**, we need to be able to compute $\Delta J_\delta / \Delta F_\delta$, for any improvement in fairness δ as the threshold of problem P_δ (2). One straightforward way of doing this is to solve P_δ and compute each change term ΔJ_δ and ΔF_δ. The problem with this approach is that we do not know a priori what the price of fairness is before solving P_δ. Given that this optimization problem can be hard-to-solve (or even not possible to solve), this approach is not efficient (or even feasible) in practice. In some situation, one may be able to compute the exact optimal of P_δ, and thus get the *tradeoff* directly by computing the solutions, but those are not the most common cases. Hence, we are going to explore ways of computing this tradeoff or approximations of it for different cases in the sections that follow.

2 Relevant Measures

2.1 Approximation Error Measures

Three common approximation errors for these kinds of models are Mean Squared Error (MSE), Mean Absolute Error (MAE), and Huber Loss. We are mainly going to focus on models that minimize the MSE, for simplicity for our analysis and because of its properties, i.e., our approximation error function would be

$$J(\beta) := \frac{1}{n} \sum_{i=1}^{n} (y_i - f_\theta(x_i))^2,$$

where $\{(x_i, y_i)\}_{i=1}^n$ is a given dataset of n samples that lie in $\mathcal{X} \times \mathcal{Y}$.

2.2 Fairness Level Measures

We explore different alternatives for fairness measures. Let $\mathcal{G}$ be a partition of the indices $I = \{1, \ldots, n\}$ that represents the split of the data according to some sensible attribute (e.g., data grouped by gender, race, or even each sample by themselves). Let $J_g(\theta)$ be the approximation error of the group $g \in \mathcal{G}$, i.e., $J_g(\theta) = \frac{1}{|g|} \sum_{i \in g} (y_i - f_\theta(x_i))^2$, then we define the following metrics to measure the level of unfairness.

1. **Maximum approximation error:** The maximum level of approximation error that a group $g \in \mathcal{G}$ achieves in the solution θ.
$$F(\theta) = \max_{g \in \mathcal{G}} J_g(\theta) \tag{5}$$

2. **Maximum approximation error difference:** The maximum distance on the level of approximation error between two groups $g, g' \in \mathcal{G}$ in the solution θ.
$$F(\theta) = \max_{g \in \mathcal{G}} |J_g(\theta) - J_{g'}(\theta)| \tag{6}$$

3. **Maximum approximation error squared difference:** The maximum distance on the level of approximation error between two groups $g, g' \in \mathcal{G}$ in the solution θ.

$$F(\theta) = \max_{g \in \mathcal{G}} (J_g(\theta) - J_{g'}(\theta))^2 \tag{7}$$

Further metrics can be considered, but our analysis will focus on these commonly used ones.

3 Approximation Models

3.1 Linear Approximation Models

In this section, we explore the hypothesis space of the linear models, which depend on the parameters $\beta \in \mathbb{R}^p$, i.e., the approximation function is defined as $f_\beta(x) = x^\top \beta$, where $x \in \mathbb{R}^p$. In this case, the approximation error function is

$$J(\beta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - x_i^\top \beta)^2 = \frac{1}{n} (y - X^\top \beta)^\top (y - X^\top \beta).$$

These models have a closed-form solution for β_0, when the matrix $X^\top X$ is invertible, so we don't even need to solve the original problem P_0 (1) in this case. This is not necessarily the case when we add the fairness constraints and try to solve the problem P_δ (2).

3.1.1 Tradeoff between approximation error and fairness

Note that, in this case, our approximation error is quadratic and twice differentiable. Let $v_0 = \nabla J(\beta_0) = 0$ (first-order condition) and $H_0 = \nabla^2 J(\beta_0)$ be its gradient and the Hessian at β_0, respectively, and let $\|u\|_A = u^\top A u$ be the norm induced by some positive semidefinite matrix $A \succeq 0$. Since J is a quadratic function, we have the exact expression for changes in it from the second-order Taylor expansion, i.e., $J(\beta_0 + \Delta\beta) - J(\beta_0) = v_0^\top \Delta\beta + \frac{1}{2} \|\Delta\beta\|_{H_0}^2 = \frac{1}{2} \|\Delta\beta\|_{H_0}^2$, for all $\Delta\beta$ in $\mathbb{R}^p$. We can then write problem P_δ (2) for this case as

$$\begin{aligned} \min_{\Delta\beta} \quad & \tfrac{1}{2} \|\Delta\beta\|_{H_0}^2 \\ s.t. \quad & F(\beta_0 + \Delta\beta) - F(\beta_0) \leq -\delta. \end{aligned} \tag{8}$$

Lower bound on the tradeoff. If $F(\beta) = \max_{g \in \mathcal{G}} J_g(\beta)$, then it is the maximum among convex functions, so it is also a convex function. The only detail to be considered is that it is not differentiable everywhere, and thus we have the convexity inequality holding with the subgradients, i.e.,

$$F(\beta_0 + \Delta\beta) \geq F(\beta_0) + a_0^\top \Delta\beta, \quad \forall \Delta\beta \in \mathbb{R}^p, \forall a_0 \in \partial F(\beta_0), \tag{9}$$

where $\partial F(\beta_0)$ is the subdifferential set of F in β_0, i.e., letting $\mathcal{A}_0 = \{g \in \arg\max_{g \in \mathcal{G}} J_g(\beta_0)\}$ be the active set of groups that achieve the maximum error, then $\partial F(\beta_0) = \text{conv}(\{\nabla J_g(\beta_0) : g \in \mathcal{A}_0\})$ is the convex combination of all the gradients of the groups in the active set $\mathcal{A}_0$.

Let $a_0 \in \partial F(\beta_0)$ be an arbitrary subgradient of F at β_0. We can now write a relaxed version of problem (8) using the lower bound for ΔF_δ using (9). Since it is a relaxed version of the fairness-constrained problem (8), it defines a lower bound on the objective, ΔJ_δ. Thus, we are interested in solving

$$\begin{aligned} \min_{\Delta\beta} \quad & \tfrac{1}{2} \|\Delta\beta\|_{H_0}^2 \\ s.t. \quad & a_0^\top \Delta\beta \leq -\delta. \end{aligned} \tag{10}$$

Let us denote the objective $g(\Delta\beta) = \frac{1}{2}\|\Delta\beta\|^2_{H_0}$ and the constraint function $h(\Delta\beta) = a_0^\top \Delta\beta + \delta$. Since g is a convex differentiable function and h is an affine function, this is a convex problem. Thus, we can use the first-order condition to get an optimal solution.

$$-\nabla g(\Delta\beta) \in \mathcal{N}_C(\Delta\beta) \iff -H_0\Delta\beta \in \{\lambda a_0 : \lambda \geq 0\},$$

where $\mathcal{N}_C(\Delta\beta)$ denotes the normal cone of $C = \{\Delta\beta : h(\Delta\beta) = 0\}$. Thus, $\Delta\beta = -\lambda H_0^{-1} a_0$, for some $\lambda \geq 0$. Since $\Delta\beta \in C$ (not useful other way), then $\lambda = \delta/(a_0^\top H_0^{-1} a_0) = \delta/\|a_0\|^2_{H_0^{-1}}$, and therefore

$$\boxed{\Delta\beta^* = -\delta H_0^{-1} a_0/\|a_0\|^2_{H_0^{-1}}} \tag{11}$$

is the optimal solution to the relaxed problem (10), for any $a_0 \in \partial F(\beta_0)$. Note that this is the optimal direction of change in the lower bound of ΔF_δ (9), for ΔJ_δ to be minimal. However, the lower bound (9) is the first-order approximation of the maximum of quadratic functions, which for an infinitesimal change in the direction $\Delta\beta_{LB}$ matches the function itself, i.e., for sufficiently small $t > 0$, $F(\beta_0 + t\Delta\beta^*) - F(\beta_0 + t\Delta\beta^*) = a_0^\top t\Delta\beta^*$ (directional derivative). Given this, for now, we can focus on changes only in the direction $\Delta\beta^*$.

Let us derive now the amount of change in the optimal cost along $\Delta\beta^*$, for a given $a_0 \in \partial F(\beta_0)$, i.e.,

$$\boxed{J(\beta_0 + \Delta\beta^*) - J(\beta_0) = \frac{1}{2}\|\Delta\beta^*\|^2_{H_0} = \delta^2/\left(2\|a_0\|^2_{H_0^{-1}}\right).} \tag{12}$$

Recalling that this is the cost change of the relaxed problem (10), then these are **lower bounds** on the actual ΔJ_δ obtained via (8). Since the constraint is setting the same level $-\delta$ for each lower bound, to get the most efficient directions we can select the subgradient $a_0 \in \partial F(\beta_0)$ that produces the least possible error difference, i.e., $\overline{a_0} = \arg\max_{a_0 \in \delta F(\beta_0)} \|a_0\|^2_{H_0^{-1}}$.

Special Case: A single group achieves the maximum error. If $|\mathcal{A}_0| = 1$, i.e., $\exists!\, g \in \mathcal{G}$ such that $F(\beta_0) = J_g(\beta_0)$, then $\partial F(\beta_0) = \{\nabla J_g(\beta_0)\}$. There is only one choice for $a_0 = \nabla J_g(\beta_0)$, to be used when defining the lower bound for ΔJ_δ.

Upper bound on the tradeoff. Similarly, we can now derive an upper bound for the change in the cost. Recall that F is the maximum of the quadratic functions J_g, which can be computed via Taylor expansion, i.e., letting v_g^0 and H_g^0 be the gradient and the Hessian of J_g at β_0, respectively, we have that

$$\begin{aligned} F(\beta_0 + \Delta\beta) - F(\beta_0) &= \max_{g\in\mathcal{G}} J_g(\beta_0 + \Delta\beta) - F(\beta_0) \\ &= \max_{g\in\mathcal{G}} J_g(\beta_0) + {v_g^0}^\top \Delta\beta + \frac{1}{2}\|\Delta\beta\|^2_{H_g^0} - F(\beta_0) \\ &\leq \max_{g\in\mathcal{G}} {v_g^0}^\top \Delta\beta + \max_{g\in\mathcal{G}} \frac{1}{2}\|\Delta\beta\|^2_{H_g^0}. \end{aligned}$$

As we did in the lower bound, we could try to solve the problem

$$\begin{aligned} \min_{\Delta\beta} \quad & \tfrac{1}{2}\|\Delta\beta\|^2_{H_0} \\ s.t. \quad & \max_{g\in\mathcal{G}} {v_g^0}^\top \Delta\beta + \max_{g\in\mathcal{G}} \tfrac{1}{2}\|\Delta\beta\|^2_{H_g^0} \leq -\delta, \end{aligned}$$

but this is not a problem with a straightforward solution. Instead, we could just choose to move in the direction $\Delta\beta^*$ (17) derived previously, and solve the problem by choosing how much we want to move in that direction, i.e., we solve the problem for $\Delta\beta = t\Delta\beta^*$;

$$\begin{aligned} \min_{t\in\mathbb{R}_{\geq 0}} \quad & \tfrac{t^2}{2}\,\|\Delta\beta^*\|^2_{H_0} \\ s.t. \quad & t\max_{g\in\mathcal{G}} {v_g^0}^\top \Delta\beta^* + \tfrac{t^2}{2}\max_{g\in\mathcal{G}}\|\Delta\beta^*\|^2_{H_g^0} \leq -\delta, \end{aligned}$$

which is a single-variable convex problem. Since the terms in the objectives are nonnegative, we can solve the quadratic equation of the constraint, with equality, and then choose the minimum nonnegative t. In other words, we get

$$t^* = \arg\min_{t\geq 0}\left\{ \frac{-\max_{g\in\mathcal{G}} {v_g^0}^\top \Delta\beta^* + s\sqrt{(\max_{g\in\mathcal{G}} {v_g^0}^\top \Delta\beta^*)^2 - 2\max_{g\in\mathcal{G}}\|\Delta\beta^*\|^2_{H_g^0}\,\delta}}{\max_{g\in\mathcal{G}}\|\Delta\beta^*\|^2_{H_g^0}} : s\in\{-1,1\}\right\},$$

which replacing in the objective, gets us something of the form

$$\boxed{J(\beta_0 + t^*\Delta\beta^*) - J(\beta_0) = \frac{t^{*2}}{2}\,\|\Delta\beta^*\|^2_{H_0} = \frac{t^{*2}}{2}\,\|a_0\|^2_{H_0^{-1}}}, \tag{13}$$

where every term of t^* is pre-computable, and we don't need to solve the optimization problem again to get this **upper bound** on ΔJ_δ.

- **Q4:** When are these bounds tight? How close are they to the actual difference in error ΔJ_δ?

[PENDING] Reason: It can be done at any time.

3.2 Generalized Linear Approximation (GLM's)

In this section, we explore the hypothesis space of the generalized linear models (GLM's), which depend on the parameters $\beta \in \mathbb{R}^p$ in the form $f_\beta(x) = g(x^\top\beta)$, where $x \in \mathbb{R}^p$, and g is a convex function. Examples include linear regression, logistic regression, and Poisson regression, among others. Defining $\ell(y_i; x_i, \beta)$ as the loss term of a given generalized linear model, the empirical risk function can be written as the average of these loss terms, i.e., we define the **empirical risk loss function** as the regularized negative log-likelihood

$$\mathcal{L}(\beta) := \frac{1}{n}\sum_{i=1}^n \ell(y_i; x_i, \beta) = \frac{1}{n}\sum_{i=1}^n (\psi(x_i^\top\beta) - y_i(x_i^\top\beta)),$$

where ψ is the function that defines the canonical-link generalized linear models that we are going to analyze, i.e., $\psi'(X^\top\beta) = \mathbb{E}[y|X;\beta]$ (e.g., $\psi(z) = z^2/2$ for linear regression). In general, these models do not have a closed-form solution for β_0 (linear regression is a special case), so we do need to at least solve the original problem P_0 (1) in these cases. Also, note that we are not focusing on the tradeoff between the loss function $\mathcal{L}$ and the (un)fairness function F, but we really want to measure the tradeoff between the empirical risk J and the (un)fairness function F. For this, we need to also define the **mean squared residual** in general form as

$$J(\beta) := \frac{1}{n}\sum_{i=1}^n r_i^2 = \frac{1}{n}\sum_{i=1}^n \Big(y_i - \psi'(x_i^\top\beta)\Big)^2,$$

where $f_\beta(x_i) = \psi'(x_i^\top \beta)$ is our approximation function for the value $y_i \in \mathcal{Y}$. This is also a convex function for the different GLMs.

Note that, in the case of the linear regression, minimizing J and minimizing $\mathcal{L}$ are equivalent optimization problems. Although, this is not generally true for GLMs. From the first-order condition of minimizing $\mathcal{L}$ we get that

$$\nabla \mathcal{L}(\beta) = \frac{1}{n} \sum_{i=1}^{n} \left(\psi'(x_i^\top \beta) - y_i \right) x_i = 0.$$

Thus, if we have a perfect approximation (i.e., $J(\beta) = 0$), the first order condition for the empirical risk is satisfied as a consequence of the zero-approximation error. The converse is not true CONTINUE...

3.2.1 Tradeoff between approximation error and fairness

In this case, the loss function is convex and twice differentiable, but it is not quadratic in general (e.g., logistic regression, Poisson regression). Nevertheless, the first-order condition still holds in the unconstrained case, so we still define $v_0 = \nabla \mathcal{L}(\beta_0) = 0$ and $H_0 = \nabla^2 \mathcal{L}(\beta_0) \succ 0$ to be its gradient and the Hessian at β_0, respectively.

Note that, with the regularization term, L is an m-strongly convex function, for some $m > 0$ that depends on the GLM, and so we have the following

$$\begin{aligned} & L(\beta_0 + \Delta\beta) - L(\beta_0) \geq v_0^\top \Delta\beta + \frac{m}{2} \|\Delta\beta\|_2^2, && \forall \Delta\beta \in \mathbb{R}^p \\ \iff & J(\beta_0 + \Delta\beta) - J(\beta_0) + \lambda \|\beta_0 + \Delta\beta\|_2^2 - \lambda \|\beta_0\|_2^2 \geq \frac{m}{2} \|\Delta\beta\|_2^2, && \forall \Delta\beta \in \mathbb{R}^p. \end{aligned}$$

Thus, we can only write a lower bound for problem P_δ (2) with this fact (using the loss function L), instead of an exact expression, but we still can recover a lower bound on ΔJ_δ from its solution, i.e, we aim to solve

$$\begin{array}{lll} \min\limits_{\Delta\beta} & \frac{m}{2} \|\Delta\beta\|_2^2 & \leq \min\limits_{\Delta\beta} \quad L(\beta_0 + \Delta\beta) - L(\beta_0) \\ s.t. & F(\beta_0 + \Delta\beta) - F(\beta_0) \leq -\delta. & s.t. \quad F(\beta_0 + \Delta\beta) - F(\beta_0) \leq -\delta. \end{array} \tag{14}$$

Lower bound on the tradeoff. Again, $F(\beta) = \max_{g \in \mathcal{G}} J_g(\beta)$ is the maximum among convex functions, thus convex. We still have that

$$F(\beta_0 + \Delta\beta) \geq F(\beta_0) + a_0^\top \Delta\beta, \quad \forall \Delta\beta \in \mathbb{R}^p, \forall a_0 \in \partial F(\beta_0). \tag{15}$$

We can then solve the relaxed version of the lower bound problem, similar as before, i.e.,

$$\begin{array}{ll} \min\limits_{\Delta\beta} & \frac{m}{2} \|\Delta\beta\|_2^2 \\ s.t. & a_0^\top \Delta\beta \leq -\delta. \end{array} \tag{16}$$

We use the same first-order condition to get an optimal solution, from $-(m\Delta\beta) = \theta a_0 \iff \Delta\beta = -\frac{\theta}{m} a_0$, for some $\theta \geq 0$. Imposing the equality on the constraint, we have $\theta 77 = m\delta / \|a_0\|_2^2$, and therefore

$$\boxed{\Delta\beta^* = -\delta a_0 / \|a_0\|_2^2} \tag{17}$$

is the optimal solution to the relaxed lower bound problem, for any $a_0 \in \partial F(\beta_0)$. Thus, a lower bound on the change in the optimal cost, for a given $a_0 \in \partial F(\beta_0)$, is

$$L(\beta_0 + \Delta\beta^*) - L(\beta_0) \geq \frac{m}{2} \|\Delta\beta^*\|_2^2$$

$$\iff J(\beta_0 + \Delta\beta^*) - J(\beta_0) \geq \frac{m\delta^2}{2\|a_0\|_2^2} - \lambda \|\beta_0 + \Delta\beta^*\|_2^2 + \lambda \|\beta_0\|_2^2 \tag{18}$$

$$\iff \boxed{J(\beta_0 + \Delta\beta^*) - J(\beta_0) \geq \frac{m\delta^2}{2\|a_0\|_2^2} - \lambda \left\| \beta_0 - \frac{\delta a_0}{\|a_0\|_2^2} \right\|_2^2 + \lambda \|\beta_0\|_2^2} \tag{19}$$

Again, since the constraint is setting the same level $-\delta$ for each lower bound, to get the most efficient direction $\Delta\beta^*$ we can select the subgradient $a_0 \in \partial F(\beta_0)$ that produces the least possible error difference, i.e., $\overline{a_0} = \arg\max_{a_0 \in \delta F(\beta_0)} \|a_0\|_2^2$.

Special Case: $J_g(\beta)$ are m_g-strongly convex. In this case, we can further try to get a tighter bound in the same direction $\Delta\beta^*$, using the m_g-strong convexity of each group's loss, i.e.,

$$F(\beta_0 + \Delta t\beta^*) - F(\beta_0) \geq t\nabla J_g(\beta_0)^\top \Delta\beta^* + \frac{m_g t^2}{2} \|\Delta\beta^*\|_2^2, \quad \forall t > 0, \forall g \in \mathcal{A}_0$$

$$\Rightarrow F(\beta_0 + \Delta t\beta^*) - F(\beta_0) \geq t{a_0}^\top \Delta\beta^* + \frac{m_0 t^2}{2} \|\Delta\beta^*\|_2^2, \quad \forall t > 0, \forall a_0 \in \partial F(\beta_0),$$

and we can define a tighter lower bound problem of the form

$$\begin{array}{ll} \min\limits_{t} & \frac{mt^2}{2} \|\Delta\beta^*\|_2^2 \\ s.t. & t a_0^\top \Delta\beta^* + \frac{m_0 t^2}{2} \|\Delta\beta^*\|_2^2 \leq -\delta, \end{array} \tag{20}$$

which we can solve setting $a_0^\top \Delta\beta^* + \frac{m_0 t^2}{2} \|\Delta\beta^*\|_2^2 + \delta = 0$, and hence $t^* = \frac{-a_0^\top \Delta\beta^* - \sqrt{(a_0^\top \Delta\beta^*)^2 - 4(\frac{m_0}{2}\|\Delta\beta^*\|_2^2)\delta}}{2(\frac{m_0}{2}\|\Delta\beta^*\|_2^2)} = \frac{\delta - \sqrt{\delta^2 - 2m_0\delta^3/\|a_0\|_2^2}}{m_0\delta^2/\|a_0\|_2^2} = \frac{1 - \sqrt{1 - 2m_0\delta/\|a_0\|_2^2}}{m_0\delta/\|a_0\|_2^2}$. Thus, using this, we get a tighter bound of the form

$$\boxed{J(\beta_0 + t^*\Delta\beta^*) - J(\beta_0) \geq \frac{m}{2} \|t^*\Delta\beta^*\|_2^2 = \frac{m\delta^2}{2\|a_0\|_2^2} \left(\frac{1 - \sqrt{1 - 2m_0\delta/\|a_0\|_2^2}}{m_0\delta/\|a_0\|_2^2} \right)^2.} \tag{21}$$

Thus, if $t^* > 1$, we clearly get a tighter lower bound compared to the previous bound on (19).

Upper bound on the tradeoff. The upper bound is trickier in this case, since every loss function behaves differently in the canonical link cases of GLM. Linear regression is a quadratic loss, Poisson is an exponential loss (not even L-smooth), etc. One question that motivates the approach that comes next is

Q: *Can we approximate the loss functions by a quadratic function or a well-behaved function in some neighborhood of β_0?*

One way a function can behave like a quadratic one in a neighborhood is by constructing an ellipsoid around the given point and analyzing the behavior of the Hessian. In other words, we are interested in knowing whether the function is **self-concordant** or not. Some common cases of the canonical link family are indeed self-concordant, but others do not. However, they satisfy a condition very similar to that of a self-concordant function, i.e., they are part of the **generalized self-concordant** function family (see Sun & Tran-Dinh (2018)), which satisfy

$$|\varphi'''(t)| \leq M_\varphi \varphi''(t)^{v/2}, \quad \forall t \in \mathbf{dom}(\varphi), \tag{22}$$

for some $v > 0$ and $M_\varphi \geq 0$. To see this, first let us consider a GLM with a canonical link $\eta_i = x_i^\top \beta, \forall i \in [n]$, then the error metric (without regularization) we are interested in minimizing is of the form

$$J(\beta) \ = \ \frac{1}{n}\sum_{i=1}^{n} \big(\psi(\eta_i) - y_i \eta_i\big), \qquad \eta_i = x_i^\top \beta,$$

where $\psi : \mathbb{R} \to \mathbb{R}$ is convex, C^3. Let us define then $\phi(t) \ := \ J(\beta_0 + td)$. We want to check (ideally prove, but pending) that it satisfies some generalized self-concordant condition. Let us check first the three most common cases.

- **Linear Regression:** In this case, $\psi(z) = \frac{z^2}{2}$, thus $\psi'(z) = z$, $\psi''(z) = 1$, and $\psi'''(z) = 0$, so it is sufficient to set $M_\psi \geq 0$ and $v > 0$.
- **Logistic Regression:** Here, $\psi(z) = log(1 + e^z)$, thus $\psi'(z) = \frac{e^z}{1+e^z}$, $\psi''(z) = \frac{e^z}{(1+e^z)^2}$, and $\psi'''(z) = \frac{e^z}{(1+e^z)^2}\frac{(1-e^z)}{(1+e^z)}$. Noting that $\frac{(1-e^z)}{(1+e^z)} \leq 1$, then it is sufficient to set $M_\psi = 1$ and $v = 2$.
- **Poisson Regression:** Now, $\psi(z) = e^z$, thus $\psi'(z) = \psi''(z) = \psi'''(z) = e^z$, and thus it is sufficient to set $M_\psi = 1$ and $v = 2$.

**Note that in the three cases we can set $v = 2$, so that $|\varphi'''(t)| \leq M_\varphi \varphi''(t), \forall t \in \mathbf{dom}(\varphi)$*

Let us assume now that $\psi(z)$ is a generalized self-concordant function with $M_\psi \geq 0$, and $v = 2$. It follows that, letting $\eta_i(t) = x_i^\top(\beta_0 + td)$,

$$\phi'(t) = \frac{1}{n}\sum_{i=1}^{n} \psi'(\eta_i(t))(x_i^\top d) - y_i(x_i^\top d), \quad \phi''(t) = \frac{1}{n}\sum_{i=1}^{n} \psi''(\eta_i(t))(x_i^\top d)^2, \quad \phi'''(t) = \frac{1}{n}\sum_{i=1}^{n} \psi'''(\eta_i(t))(x_i^\top d)^3,$$

and thus we can get the self-concordant condition for $\phi(t)$ using ψ as,

$$\begin{aligned}
|\phi'''(t)| &= \frac{1}{n}\left|\sum_{i=1}^{n}\psi'''(\eta_i(t))(x_i^\top d)^3\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left|\psi'''(\eta_i(t))(x_i^\top d)^3\right| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\left|\psi'''(\eta_i(t))\right|(x_i^\top d)^2|x_i^\top d| \\
&\leq \max_{j\in[n]}|x_j^\top d|\frac{1}{n}\sum_{i=1}^{n}\left|\psi'''(\eta_i(t))\right|(x_i^\top d)^2 \\
&\leq M_\psi \max_{j\in[n]}|x_j^\top d|\frac{1}{n}\sum_{i=1}^{n}\psi''(\eta_i(t))(x_i^\top d)^2 \qquad \text{(self-concordance)} \\
&= M_\phi \phi''(t),
\end{aligned}$$

and thus $\phi(t)$ is also a generalized self-concordant function with the same constant (in terms of t) $M_\phi = M_\psi \max_{j\in[n]}|x_j^\top d| \geq 0$ and $v = 2$. Now, we can focus on trying to integrate this differential inequation using the auxiliary function $u(t) = \phi''(t)$, i.e.,

$$\begin{aligned}
-M_\phi u(t) \leq & u'(t) \leq M_\phi u(t), \forall t > 0 \\
u(0)e^{-M_\phi t} \leq & u(t) \leq u(0)e^{M_\phi t} \quad \text{(Grönwall's inequality)} \\
u(0)\int_0^t e^{-M_\phi s}ds \leq & \int_0^t u(s)ds \leq u(0)\int_0^t e^{M_\phi s}ds \\
\phi'(0) - u(0)\frac{e^{-M_\phi t}-1}{M_\phi} \leq & \phi'(t) \leq \phi'(0) + u(0)\frac{e^{M_\phi t}-1}{M_\phi} \qquad \Big/\int_0^t \cdot\, ds \\
\phi'(0)t + u(0)\frac{e^{-M_\phi t}-1+M_\phi t}{M_\phi^2} \leq & \phi(t) - \phi(0) \leq \phi'(0)t + u(0)\frac{e^{M_\phi t}-1-M_\phi t}{M_\phi^2} \\
t\nabla J(\beta)^\top d + \frac{d'\nabla^2 J(\beta)d}{M_\phi^2}\left(e^{-M_\phi t}-1+M_\phi t\right) \leq & J(\beta+td) - J(\beta) \leq t\nabla J(\beta)^\top d + \frac{d'\nabla^2 J(\beta)d}{M_\phi^2}\left(e^{M_\phi t}-1-M_\phi t\right).
\end{aligned}$$

Remembering that $\nabla J(\beta_0) = 0$, we get two bounds (lower and upper) on ΔJ_δ, of the form

$$\boxed{J(\beta_0 + t\Delta\beta) - J(\beta_0) \;\geq\; \frac{\|\Delta\beta\|_{H_0}^2}{M_\phi^2}\left(e^{-M_\phi t} + M_\phi t - 1\right)} \tag{23}$$

$$\boxed{J(\beta_0 + t\Delta\beta) - J(\beta_0) \;\leq\; \frac{\|\Delta\beta\|_{H_0}^2}{M_\phi^2}\left(e^{M_\phi t} - M_\phi t - 1\right)} \tag{24}$$

- **Linear Regression:** In this case, we can choose $M_\psi = 0 \Rightarrow M_\phi = 0$, and (in the limit) recover the quadratic difference equality: $J(\beta_0 + t\Delta\beta) - J(\beta_0) = \frac{t^2}{2}\|\Delta\beta\|_{H_0}^2$.
- **Linear Regression/Logistic Regression/Poisson Regression:** We can set $M_\phi = \max_{j\in[n]}\{|x_j^\top \Delta\beta|\}$, and use the bounds above.

Note: It depends on a exponential over t, so it may be a very loose bound. Since we are working with a family of models, this could be good enough anyway. We can also apply this bound on each $g \in \mathcal{G}$, and use the same direction $\Delta\beta^*$ from a lower bound. I'm going to run some numerical experiments after deriving it properly and verify that it is correct.

3.2.2 Bounds on the *tradeoff* between J and F (independent of strong convexity)

Again, to get an approximately good direction $\Delta\beta^*$ to work with in the first place, we aim for the solution of the Taylor approximation (1st order for F and 2nd order for J) problem $\min_{\Delta\beta}\left\{\frac{1}{2}\left\|\Delta\beta\right\|_{H_0}^2 \ : \ a_0^\top\Delta\beta \leq -\delta\right\}$, where $a_0 \in \partial F(\beta_0)$. As in (ref.), we get the direction $\Delta\beta^* = -\delta H_0^{-1}a_0/\left\|a_0\right\|_{H_0^{-1}}^2$; which we can now use to get the valid bounds from (23) and (24).

- **Lower bound on the *tradeoff*.** We use (23) to lower bound both ΔJ_δ and ΔF_δ. Let us define $C_{\phi,I}^{LB}(t,d) = (e^{-tM_\psi \max_{i\in I}|x_i'd|} + t(M_\psi \max_{i\in I}|x_i'd|) - 1)/(M_\psi \max_{i\in I}|x_i'd|)^2$. We are interested in solving the lower bound problem

$$\min_{t\geq 0}\left\{C_{\phi,[n]}^{LB}(t,\Delta\beta^*)\|\Delta\beta^*\|_{H_0}^2 \ : \ t(a_{\partial F_0}^\top\Delta\beta^*) + C_{\phi,\bar{g}}^{LB}(t,\Delta\beta^*)\|\Delta\beta^*\|_{H_{\partial^2 F_0}}^2 \leq -\delta\right\}, \tag{25}$$

  where $a_{\partial F_0} \in \partial F(\beta_0)$, $H_{\delta^2 F_0} \in \partial^2 F(\beta_0)$ (second-order subdifferential), and $\bar{g}$ represents the combination of groups $g \in \mathcal{G}$ that leads to both $a_{\partial F_0}$ and $H_{\partial^2 F_0}$, so that we get the proper M_ϕ. Although this problem is very dense in notation, solving it is just finding the roots of the exponential equation of the constraint, and choosing the one that produces the minimum cost.

Upper bound (independent of L-smoothness)

Again, we can just see how much we need to advance in the direction $\Delta\beta^*$, by solving the problem $\min_{t\geq 0}\left\{C_\phi^{UB}t^2\left\|\Delta\beta^*\right\|_{H_0}^2 \ : \ C_\phi^{UB_g}t^2\left\|\Delta\beta^*\right\|_{H_0}^2 \leq -\delta\right\}$, where $C_\phi^{UB} = (e^{M_\phi} - M_\phi - 1)/M_\phi^2$, and similar for each group, b2ut

$$\boxed{J(\beta_0 + \Delta\beta^*) - J(\beta_0) \leq \frac{C_\phi^{UB}\delta^2}{\left\|a_0\right\|_{H_0^{-1}}^2}} \tag{26}$$

3.3 Redefining the (un)fairness level

We start this section by noting something we were initially ignoring. That is, the fact that, for general GLMs, the loss function does not necessarily represent the approximation error. The approximation function for a general GLM is the expected value of the assumed variable $y \in \mathcal{Y}$, given some $x \in \mathcal{X}$ and the fitted parameter β, i.e.,

$$y_i \approx \mathbb{E}[y_i|x_i;\beta] = \mu(x_i^\top\beta) = \psi'(x_i^\top\beta), \quad \forall i \in [n].$$

From this expression, we can define some approximation error as a function of the distance between the data vector y and the approximation function. As a simile to the approximation error of linear regression (MSE), we are going to work in this case with the sum of squared residuals $r_i(\beta) := \psi'(x_i^\top\beta) - y_i$, i.e.,

$$R(\beta) = \frac{1}{n}\sum_{i=1}^n r_i(\beta)^2 = \frac{1}{n}\sum_{i=1}^n\left(\psi'(x_i^\top\beta) - y_i\right)^2,$$

which is not a straightforward transformation of the loss function J, and thus cannot be represented via the group version of J. Furthermore, this function is not even convex in general. Note that the same function can be defined for every partition of indices, and so we define $R_g(\beta) := \frac{1}{n_g}\sum_{i\in g} r_i(\beta)^2, \forall g \in \mathcal{G}$. From this, we can also define the updated (nonconvex) version of the fairness function as

$$F(\beta) = \max_{g\in\mathcal{G}} R_g(\beta),$$

which raises the following natural question for tradeoff purposes.

CORRECTION: Instead of using the nonconvex squared residuals, we should instead use the proper convex measure of distance between our approximation $\psi'(x^\top\beta)$ and the true y. The loss functions of GLMs are just the Bregman Divergence in terms of their respective link. In other words, we could rewrite the objective loss functions directly using the Bregman Divergence average. This is a nonnegative discrepancy measure, and we will see that it is zero when the approximation is perfect. That is

$$J(\beta) = \frac{1}{n}\sum_{i=1}^n D_\psi(\eta_{y_i}, x_i^\top\beta) = \frac{1}{n}\sum_{i=1}^n \Big(\psi(x_i^\top\beta) - \psi(\eta_{y_i}) - \psi'(\eta_{y_i})(x_i^\top\beta - \eta_{y_i})\Big), \quad \text{where } \eta_{y_i} = \psi'^{-1}(y_i).$$

This is a convex function over β, since ψ is convex and the rest of the terms are linear on β. From this, we can define and recover the convexity of our (un)fairness measure, just letting each group partition of indices have their own term defined in the same way as originally discussed.

Furthermore, let us observe the gradient and Hessian of this loss function.

$$\nabla J(\beta) = \frac{1}{n}\sum_{i=1}^n \Big(\psi'(x_i^\top\beta) - \psi'(\eta_{y_i})\Big)x_i = \frac{1}{n}\sum_{i=1}^n \Big(\psi'(x_i^\top\beta) - y_i\Big)x_i,$$

and thus note that we satisfy the first order condition if $\psi'(x_i^\top\beta) = y_i, \forall i \in [n]$. Further, with this condition, the loss function becomes zero. The Hessian is the same as in any loss of GLMs, i.e,

$$\nabla^2 J(\beta) = \frac{1}{n}\sum_{i=1}^n \psi''(x_i^\top\beta)x_i x_i^\top = \frac{1}{n}X^\top \mathbf{diag}\Big(\psi''(x_1^\top\beta), \ldots, \psi''(x_n^\top\beta)\Big)X,$$

Hence, we recover a convex loss function and a convex fairness metric, since we define it as the maximum of these convex function, i.e, $F(\beta) = \max_{g\in\mathcal{G}}\{J_g(\beta)\}$.

- **Linear Regression:** We recover an MSE transformation, given that $\psi(z) = \frac{z^2}{2}$, and so $D_\psi(\eta_i, x_i^\top\beta) = \frac{1}{2}(x_i^\top\beta - y_i)^2 \Rightarrow J(\beta) = \frac{1}{2n}\sum_{i=1}^n (x_i^\top\beta - y_i)^2$.
- **Logistic Regression:** We recover the cross-entropy, since $\psi(z) = \log(1+e^z)$, and also from the Bregman divergence relationship with the convex conjugate we have that $D_\psi(\eta_i, x_i^\top\beta) = D_{\psi^*}(\psi'(x_i^\top\beta), y_i) = -\psi'(x_i^\top\beta)\log(y_i) - (1-\psi'(x_i^\top\beta))\log(1-y_i)$.
- **Poisson Regression:**

Q: *Can we analyze or derive bounds for the tradeoff between J and F in this general setting for GLMs?*

To answer this question, we first make use of the definitions of relative strong convexity and relative smoothness described in Lu, Freund, & Nesterov (2017). Let $h(\cdot)$ be any differentiable convex function on some set Q, which we call the "reference function". We have the following two definitions.

- **Relative Smoothness:** $f(\cdot)$ is L-smooth relative to $h(\cdot)$ on $Q \iff \forall x, y \in \mathbf{int}(Q),\ \exists L > 0$, such that
$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + L D_h(y, x).$$
- **Relative Smoothness:** $f(\cdot)$ is μ-strongly convex relative to $h(\cdot)$ on $Q \iff \forall x, y \in \mathbf{int}(Q),\ \exists \mu \geq 0$, such that
$$f(y) \geq f(x) + \langle \nabla f(x), y - x \rangle + \mu D_h(y, x).$$

If $f(\cdot)$ is twice differentiable, then we can also make use of the equivalent definitions

$$\mu \nabla^2 h(x) \preceq \nabla f(x) \preceq L \nabla^2 h(x), \quad \forall x, y \in \mathbf{int}(Q).$$

Relative strong convexity of F. Let us derive the Hessian matrix for $J(\beta)$ and $R(\beta)$, i.e.

$$\begin{aligned}
\nabla^2 J(\beta) &= \frac{1}{n} \sum_{i=1}^{n} \psi''(x_i^\top \beta) x_i x_i^\top \\
&= \frac{1}{n} X^\top \mathbf{diag}\left(\psi''(x_1^\top \beta), \ldots, \psi''(x_n^\top \beta) \right) X,
\end{aligned}$$

$$\begin{aligned}
\nabla^2 R(\beta) &= \frac{1}{n} \sum_{i=1}^{n} \nabla^2 \big(\psi'(x_i^\top \beta) - y_i \big)^2 \\
&= \frac{2}{n} \sum_{i=1}^{n} \left(\psi''(x_i^\top \beta)^2 + (\psi'(x_i^\top \beta) - y_i) \psi'''(x_i^\top \beta) \right) x_i x_i^\top \\
&= \frac{2}{n} X^\top \mathbf{diag}\left(\psi''(x_i^\top \beta)^2 + (\psi'(x_i^\top \beta) - y_i) \psi'''(x_i^\top \beta) \right) X.
\end{aligned}$$

Thus, if we had that $\frac{L}{2} \psi''(x_i^\top \beta) \geq \psi''(x_i^\top \beta)^2 + (\psi'(x_i^\top \beta) - y_i) \psi'''(x_i^\top \beta) \geq \frac{\mu}{2} \psi''(x_i^\top \beta), \forall i \in [n]$, for some $\mu \geq 0$ and $L > 0$, we would have relatively μ-strong convexity and relatively L-smoothness. We can analyze further this term, assuming that $\exists M \gg 0$, sufficiently large such that $x_i^\top \beta \in [-M, M]$, for all practical solutions $\beta \in \mathbb{R}^p$. Hence, $\psi''(x_i^\top \beta) > 0$ for all β of interest. We then rewrite the conditions as

$$\begin{aligned}
&\frac{L}{2} \psi''(x_i^\top \beta) \geq \psi''(x_i^\top \beta)^2 + (\psi'(x_i^\top \beta) - y_i) \psi'''(x_i^\top \beta) \geq \frac{\mu}{2} \psi''(x_i^\top \beta), && \forall i \in [n] \\
\iff\ &\frac{L}{2} \geq \psi''(x_i^\top \beta) + (\psi'(x_i^\top \beta) - y_i) \frac{\psi'''(x_i^\top \beta)}{\psi''(x_i^\top \beta)} \geq \frac{\mu}{2}, && \forall i \in [n].
\end{aligned}$$

- **Linear Regression:** In this case, $\psi''(z) = 1 \ \wedge \ \psi'''(z) = 0$, and thus we can just choose any $L \geq 2$ and $\mu \leq 2$, and we recover the standard definitions. Note that we don't even need the assumption.

- **Logistic Regression:** Here, note that $\psi'(z) \in [0,1]$ and $y \in [0,1]$, so $r_i \in [-1,1]$, and also $\frac{\psi'''(z)}{\psi''(z)} = \frac{(1-e^z)}{(1+e^z)} \in [-1,1]$. Further $\psi''(z) \leq \frac{1}{4}$. Thus, for **relative smoothness**, we only need to check
$$L \geq 1/2 + 2 = \frac{5}{2}.$$
Note now that $\psi''(z) = \psi'(z)/(1+e^z)$, and therefore we can rewrite the condition for **relative strong convexity** as
$$\frac{\psi'(x_i^\top \beta)}{(1+e^{x_i^\top \beta})} + (\psi'(x_i^\top \beta) - y_i)\frac{(1-e^{x_i^\top \beta})}{(1+e^{x_i^\top \beta})} \geq \frac{\mu}{2}.$$
Since $x_i^\top \beta \in [-M, M]$, for some $M \gg 0$, then $\frac{1}{4} \geq \psi''(x_i^\top \beta) \geq \varepsilon_2$, for some $1 \gg \varepsilon_2 > 0$ ($\varepsilon_2 \approx 0$), and $\varepsilon_1 \leq \psi'(x_i^\top \beta) \leq 1 - \varepsilon_1$, for some $\varepsilon_1 > 0$ ($\varepsilon_1 \approx 0$).

- **Poisson Regression:** In this case, $\psi(z) = \psi'(z) = \psi''(z) = \psi'''(z)$, so

Note that the relatively L-smoothness is straightforward, since

Blank

For each group $g \in \mathcal{G}$, we define the same error metric $J_g(\beta) = \frac{1}{n_g}\sum_{i\in g}\big(\psi(\eta_i) - y_i\eta_i\big)$, and, again, the maximum of them is the fairness $F(\beta) = \max_{g\in\mathcal{G}} J_g(\beta)$.. Let us forget about the regularization part to begin with, and redefine

$$H_0 \;=\; \nabla^2 J(\beta_0) = \frac{1}{n} X^\top W_0 X, \qquad H_{0,g} \;=\; \nabla^2 J_g(\beta_0) = \frac{1}{n_g} X_g^\top W_{0,g} X_g,$$

with $W_0 = \mathrm{diag}(\psi''(\eta_i(\beta_0)))$, and similarly $W_{0,g}$ restricted to group g.

3.4 Generalized self–concordance (Bach-type) along a ray

Assume the *generalized SC* condition:

$$|\psi'''(z)| \;\le\; M\,\psi''(z) \qquad \text{for all } z \in \mathbb{R}, \tag{27}$$

with a model-dependent constant $M \ge 0$ (Gaussian: $M = 0$; logistic/Poisson: $M = 1$).

Fix a direction $d \in \mathbb{R}^p$ and define the 1D restriction

$$\phi(t) \;:=\; J(\beta_0 + td).$$

Let

$$Q \;:=\; d^\top H_0 d \;=\; \phi''(0), \qquad \rho \;:=\; M\,\max_i |x_i^\top d| \;>0.$$

Lemma (ray-wise envelope for J). Under (27), for all $t \ge 0$,

$$J(\beta_0 + td) \;\le\; J(\beta_0) + \nabla J(\beta_0)^\top (td) \;+\; \frac{Q}{\rho^2}\Big(e^{\rho t} - \rho t - 1\Big). \tag{28}$$

Proof sketch. Set $y(t) = \phi''(t)$. Using (27) and $\zeta_i := x_i^\top d$, $|\phi'''(t)| \le \sum_i |\psi'''(\eta_i(t))|\,|\zeta_i|^3/n \le M(\max_i|\zeta_i|)\sum_i \psi''(\eta_i(t))\zeta_i^2/n = \rho\,\phi''(t)$. Thus $-\rho y \le y' \le \rho y$, and Grönwall implies $y(0)e^{-\rho t} \le y(t) \le y(0)e^{\rho t}$. Integrate twice: $\phi'(t) \le \phi'(0) + y(0)(e^{\rho t}-1)/\rho$ and $\phi(t) \le \phi(0) + \phi'(0)t + y(0)(e^{\rho t} - \rho t - 1)/\rho^2$. Since $y(0) = Q$ and $\phi'(0) = \nabla J(\beta_0)^\top d$, we get (28). □

Groupwise envelopes. For each group g, define

$$A_g \;:=\; \nabla J_g(\beta_0)^\top d, \qquad Q_g \;:=\; d^\top H_{0,g} d, \qquad \rho_g \;:=\; M_g \max_{i\in g} |x_i^\top d|,$$

where M_g satisfies $|\psi_g'''| \le M_g \psi_g''$ for that group/model. Then for all $t \ge 0$,

$$J_g(\beta_0 - td) \;\le\; J_g(\beta_0) \;-\; A_g t \;+\; \frac{Q_g}{\rho_g^2}\Big(e^{\rho_g t} - \rho_g t - 1\Big). \tag{29}$$

3.5 Feasible step for a fairness tightening δ

Fix a target $\delta > 0$ and a direction d. Define the groupwise convex functions

$$f_g(t) \;:=\; -\,A_g\,t \;+\; \frac{Q_g}{\rho_g^2}\Big(e^{\rho_g t} - \rho_g t - 1\Big), \qquad t \ge 0. \tag{30}$$

They satisfy

$$f_g(0) = 0, \qquad f_g'(t) = -A_g + \frac{Q_g}{\rho_g}\big(e^{\rho_g t} - 1\big), \qquad f_g''(t) = Q_g e^{\rho_g t} > 0,$$

hence each f_g is strictly convex and has a unique minimizer at

$$t_{0,g} = \frac{1}{\rho_g} \log\Big(1 + \frac{A_g \rho_g}{Q_g}\Big) \quad \text{(provided } A_g > 0\text{)}. \tag{31}$$

Using (29), the fairness constraint

$$F(\beta_0 - td) = \max_g J_g(\beta_0 - td) \;\le\; F(\beta_0) - \delta$$

is *sufficiently* enforced if and only if

$$f_g(t) \;\le\; -\delta \qquad \text{for all groups } g. \tag{32}$$

For each g, let $t_g(\delta)$ be the smallest root of $f_g(t) = -\delta$ (which lies in $[0, t_{0,g}]$ whenever feasible). Then the *minimal feasible step along* d is

$$t^\star(d, \delta) \;=\; \max_g t_g(\delta). \tag{33}$$

Practical computation of $t_g(\delta)$. Define $h_g(t) := f_g(t) + \delta$. Then $h_g(0) = \delta > 0$, $h_g'(0) = -A_g < 0$, and h_g is strictly convex. A safe upper bracket follows from $e^x \ge 1 + x + \frac{1}{2}x^2$:

$$h_g(t) \;\ge\; \frac{1}{2} Q_g t^2 - A_g t + \delta,$$

whose smallest positive root is

$$t_g^{\text{quad}}(\delta) \;=\; \frac{A_g - \sqrt{A_g^2 - 2Q_g\delta}}{Q_g}. \tag{34}$$

Since $h_g \ge$ (quadratic), the true root satisfies $t_g(\delta) \le t_g^{\text{quad}}(\delta)$. A (projected) Newton step converges rapidly:

$$t \;\leftarrow\; \Pi_{[0,\, t_g^{\text{quad}}]}\Big(t - \frac{h_g(t)}{h_g'(t)}\Big), \qquad h_g'(t) = -A_g + \frac{Q_g}{\rho_g}\big(e^{\rho_g t} - 1\big).$$

For very small δ, the initialization $t_g^{(0)} = \delta / A_g$ is excellent, and the second-order expansion is

$$t_g(\delta) \;=\; \frac{\delta}{A_g} + \frac{Q_g}{2A_g^3}\,\delta^2 + O(\delta^3). \tag{35}$$

3.6 Risk upper bound once $t^\star$ is known

Apply the envelope (28) with $Q = d^\top H_0 d$ and $\rho = M \max_i |x_i^\top d|$:

$$\Delta J \;:=\; J(\beta_0 - t^\star d) - J(\beta_0) \;\le\; \nabla J(\beta_0)^\top(-t^\star d) \;+\; \frac{Q}{\rho^2}\Big(e^{\rho t^\star} - \rho t^\star - 1\Big). \tag{36}$$

If β_0 is a stationary point of J (e.g., the unconstrained GLM fit), the linear term vanishes:

$$\boxed{\Delta J \;\le\; \frac{Q}{\rho^2}\Big(e^{\rho t^\star} - \rho t^\star - 1\Big).} \tag{37}$$

Small–δ expansion (consistency with quadratic theory). From (35), $t^\star = \max_g t_g(\delta) = \frac{\delta}{A_{\min}} + O(\delta^2)$ with $A_{\min}(d) := \min_g A_g$. Using $e^{\rho t} - \rho t - 1 = \frac{1}{2}(\rho t)^2 + O(t^3)$,

$$\Delta J \;\le\; \frac{Q}{2}\,(t^\star)^2 + O((t^\star)^3) \;=\; \frac{1}{2}\frac{Q}{A_{\min}^2}\,\delta^2 + O(\delta^3).$$

If d is chosen as the optimal local direction $d = H_0^{-1}a^\star$ with

$$a^\star \in \arg\min_{a \in \partial F(\beta_0)} a^\top H_0^{-1} a, \qquad \partial F(\beta_0) = \mathrm{co}\{\nabla J_g(\beta_0) : g \in \arg\max J_g(\beta_0)\},$$

then $Q = d^\top H_0 d = a^{\star\top} H_0^{-1} a^\star$ and $A_{\min}(d) = a^{\star\top} d = Q$, so the leading coefficient equals the sharp lower-bound constant:

$$\Delta J \;\le\; \frac{\delta^2}{2\,a^{\star\top} H_0^{-1} a^\star} \;+\; O(\delta^3), \qquad \Delta J \;\ge\; \frac{\delta^2}{2\,a^{\star\top} H_0^{-1} a^\star} \;+\; o(\delta^2).$$

3.7 Special cases and remarks

- **Gaussian (OLS).** $M = 0 \Rightarrow \rho = 0$ and (29) reduces to $f_g(t) = -A_g t + \frac{1}{2} Q_g t^2$, giving the closed-form $t_g(\delta) = \dfrac{A_g - \sqrt{A_g^2 - 2Q_g\delta}}{Q_g}$ and $\Delta J = \frac{1}{2} Q (t^\star)^2$.
- **Logistic / Poisson.** $M = 1$; take $\rho_g = \max_{i \in g} |x_i^\top d|$, $\rho = \max_i |x_i^\top d|$.
- **Ridge GLM.** If $J_\gamma = J + \frac{\gamma}{2}\|\beta\|^2$, then replace H_0 by $H_0 + \gamma I$ in Q, Q_g, which improves curvature and makes global LBs possible.
- **Feasibility test.** For each g, feasibility along d requires $\delta \le -f_g(t_{0,g})$, with $t_{0,g}$ as in (31); if violated for some g, no $t \ge 0$ can satisfy (32) along that fixed d.

3.8 Algorithmic summary (one line per step)

1. Choose a direction d (e.g., $d = H_0^{-1}a^\star$ for the optimal local tradeoff).
2. For each group g, compute A_g, Q_g, ρ_g and solve $f_g(t) = -\delta$ using Newton with the safe bracket $[0,\, t_g^{\mathrm{quad}}(\delta)]$ in (34).
3. Set $t^\star = \max_g t_g(\delta)$ and evaluate the risk UB via (37).

Blank

Group residual means. For each group g, let

$$\overline{r}_g(t) = \frac{1}{|G_g|} \sum_{i \in G_g} r_i(t) = \frac{1}{|G_g|} \sum_{i \in G_g} \big(f(x_i; t) - y_i\big).$$

Pairwise mean residual gap. Fairness is measured by the maximum difference in mean residuals:

$$h(t) = \max_{g,g'} \left|\overline{r}_g(t) - \overline{r}_{g'}(t)\right| = \left\|KS(\hat{y}(t) - y)\right\|_\infty, \tag{38}$$

where $S \in \mathbb{R}^{G \times n}$ computes group averages, and $K \in \mathbb{R}^{m \times G}$ (with $m = \frac{1}{2}G(G-1)$) encodes all signed pairwise differences. For example, each row of K equals $e_g - e_{g'}$ for some (g, g') pair.

Optimization problem. We minimize prediction error subject to this fairness penalty:

$$\min_{t \in \mathbb{R}^p} \quad F_\lambda(t) = g(t) + \lambda\, h(t) \quad \text{with} \quad g(t) = \frac{1}{n} \sum_{i=1}^{n} \ell(f(x_i; t), y_i), \tag{39}$$

where ℓ is differentiable (e.g. squared loss) and $\lambda \geq 0$.

4 Structure and Algorithmic Regime

Equation (39) takes the composite form

$$F_\lambda(t) = g(t) + \lambda\, h(c(t)), \qquad c(t) = KS(\hat{y}(t) - y).$$

Here g is smooth and $h(u) = \|u\|_\infty$ is convex but nonsmooth. This is *Case B* (output-composite): the fairness penalty acts on a smooth transformation of parameters. We employ the **prox-linear** method with a **primal–dual inner solver**.

4.1 Linearization and Inner Subproblem

At iterate t^k, linearize $c(t)$:

$$c(t) \approx c(t^k) + J_c(t^k)(t - t^k),$$

where $J_c(t^k) = KSJ_{\hat{y}}(t^k)$ is the Jacobian of the residual aggregation. Define $q^k = t^k - \alpha \nabla g(t^k)$ and form

$$t^{k+1} \in \arg\min_t \left\{ \frac{1}{2\alpha} \left\| t - q^k \right\|^2 + \lambda \left\| A^k t + b^k \right\|_\infty \right\}, \quad A^k = J_c(t^k),\ b^k = c(t^k) - A^k t^k. \tag{40}$$

This subproblem is convex and quadratic-plus-ℓ_∞. Its dual form yields simple proximal operators for the conjugate $h^*(y) = \iota_{\{\|y\|_1 \leq 1\}}(y)$, enabling efficient *Chambolle–Pock* iterations.

4.2 Inner Primal–Dual Steps

Let $\sigma, \tau > 0$ satisfy $\tau\sigma \left\|A^k\right\|^2 < 1$. The updates are

$$\begin{aligned}
y^{\ell+1} &= \mathrm{proj}_{\|\cdot\|_1 \leq \lambda} \left(y^\ell + \sigma(A^k \bar{t}^\ell + b^k)\right), \\
t^{\ell+1} &= \frac{1}{1 + \tau/\alpha} \Big(t^\ell - \tau (A^k)^\top y^{\ell+1} + (\tau/\alpha) q^k\Big), \\
\bar{t}^{\ell+1} &= t^{\ell+1} + \theta(t^{\ell+1} - t^\ell).
\end{aligned}$$

The projection onto the ℓ_1-ball is standard: sort, threshold, and rescale.

4.3 Outer Loop

Repeat the prox-linear step with backtracking until the proximal gradient mapping $\frac{1}{\alpha}\left\|t^{k+1}-t^k\right\|$ falls below a tolerance. Under convexity, convergence is guaranteed; if g is μ-strongly convex, rates are linear.

5 Fairness–Accuracy Trade-off Geometry

Define the penalized and constrained value functions:

$$\psi(\lambda)=\min_t g(t)+\lambda h(t), \qquad \phi(\tau)=\min_t g(t) \text{ s.t. } h(t)\le\tau.$$

Proposition 1 (Envelope relationships)**.**

$$\psi'(\lambda)=h(t_\lambda), \qquad \textit{(attained fairness level)} \tag{41}$$
$$\phi'(\tau)=-\lambda^\star_\tau, \qquad \textit{(optimal Lagrange multiplier)} \tag{42}$$

whenever derivatives exist. The quantities are subgradients otherwise.

Hence the slope of ψ equals the fairness at the solution, while the negative slope of ϕ equals the dual multiplier, representing the marginal increase in prediction error per unit of fairness relaxation.

5.1 Lipschitz Sensitivity

If g is μ-strongly convex and h is G-Lipschitz ($G=1$ here), then

$$\|t_\lambda-t_{\lambda'}\|\le\frac{M}{\mu}|\lambda-\lambda'|, \qquad |h(t_\lambda)-h(t_{\lambda'})|\le\frac{M^2}{\mu}|\lambda-\lambda'|,$$

where $M=\sup_t\|J_c(t)\|$ (controlled by model Jacobian and group averaging). These bounds quantify how solutions and fairness change with λ.

6 Implementation for Regression Models

1. **Group operators.** Build S and K once from group memberships.
2. **Automatic differentiation.** Implement vector–Jacobian and Jacobian–vector products for $t\mapsto\hat{y}(t)$.
3. **Outer step size.** $\alpha\approx 1/L_g$ estimated by backtracking on g.
4. **Inner steps.** Choose τ,σ with $\tau\sigma\left\|A^k\right\|^2<1$ (estimate $\|A^k\|$ via a few power iterations).
5. **Continuation in λ.** Solve for a grid or use bisection to reach target fairness τ.
6. **Dual variable.** The inner dual y approximates the gradient of ϕ; its norm indicates the marginal cost of fairness.

7 Experimental Protocol

- **Metrics.** Report MSE(t), $h(t)$, and per-group residual means $\overline{r}_g(t)$.
- **Pareto curve.** Plot $(h(t_\lambda), g(t_\lambda))$ as λ varies. Include theoretical slope $\frac{dg^\star}{d\tau} = -\lambda_\tau^\star$ from the dual variable.
- **Certification.** Inner primal–dual gap $< \epsilon_{\text{in}}$ bounds fairness error; outer stationarity bounds error on g.

8 Discussion and Outlook

The pairwise mean residual gap fairness (38) is convex and piecewise linear, enabling exact nonsmooth optimization. Our analysis clarifies:

- The natural method is **prox-linear** (outer) with **primal–dual** (inner);
- Dual variables directly quantify the fairness–accuracy slope;
- Lipschitz bounds control sensitivity of both loss and fairness to λ;
- In nonlinear (e.g. neural) models, weak convexity suffices for convergence to critical points.

Extensions include fairness constraints on absolute residuals, robust formulations under distribution shift, and bilevel calibration of λ to match desired fairness levels.

Conclusion

We have unified modeling, algorithms, and theory for regression models regularized by the pairwise mean residual gap. This functional captures disparities in systematic over- or under-prediction across groups, and its convex yet nonsmooth structure fits seamlessly into the composite optimization framework. The resulting pipeline—prox-linear outer iterations with primal–dual inner solves— provides not only effective training but also analytical insight into how predictive accuracy and fairness interact.